AIRPORT PRICING POLICIES:
AIRLINE CONDUCT, PRICE DISCRIMINATION, DYNAMIC
CONGESTION AND NETWORK EFFECTS
This thesis is part of a European Research Council project on “Optimizing Policies for Transport: Accounting for Industrial Organization in Network Markets” (OPTION), funded under European Union’s Seventh Framework Programme (FP7/2007-2013) through Advanced Grant 246969. The ERC’s financial support is gratefully acknowledged.

ISBN 978 90 361 0429 6

Cover design: Crasborn Graphic Designers bno, Valkenburg a.d. Geul
Cover illustration: Gabriel De Ioannes

This book is no. 613 of the Tinbergen Institute Research Series, established through cooperation between Rozenberg Publishers and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.
AIRPORT PRICING POLICIES:
AIRLINE CONDUCT, PRICE DISCRIMINATION, DYNAMIC
CONGESTION AND NETWORK EFFECTS

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. F.A. van der Duyn Schouten,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Economische Wetenschappen en Bedrijfswetenschappen
op woensdag 8 april 2015 om 13.45 uur
in de aula van de universiteit,
De Boelelaan 1105

doctor

Hugo Emilio Silva Montalva

geboren te Santiago, Chili
promotor: prof.dr. E.T. Verhoef
copromotor: dr. V.A.C. van den Berg
Preface

This thesis represents much more than the research that I have carried out during the last 4 years. It is the culmination of a long process that was not only about acquiring knowledge and learning how to formulate and answer relevant research questions. It was also about learning how to be patient, about traveling, discovering different ways to see the world and being part of a different culture. It was about appreciating others, about raising a family and realizing what is important in life. It is only fair to acknowledge the ones who have been important for me during these years.

The exceptionally important and pleasant stay at the Department of Spatial Economics could have not taken place without my advisor Erik Verhoef. Thank you for giving me the opportunity to become a doctor under your supervision. A great deal of the satisfaction I feel after these 4 years are due to Erik’s high-quality advise, his flexible approach to supervision, his insightful comments, the support he gave me to develop myself, his view on economics and many more attributes that make him an outstanding academic and advisor. This is also the right place to reassert how much I enjoyed the meetings with Erik and Vincent, my co-supervisor. An outsider attending them could easily have thought that writing a paper, especially on dynamic scheduling in the bottleneck model, is about staring at the roof for five minutes without saying a word. Or that we were writing a paper on the economics of football, as most meetings featured discussions about the latest Ajax, Champions League and World Cup football results. It was during these meetings when I interacted the most with Erik and Vincent, beyond equations, models and interpretation of results, and I am very grateful for those moments. I am also grateful to Vincent for his supervision, his support and the nice discussions we had. Furthermore, I thank Erik for letting me extend my stay beyond the 4-year PhD contract. I look forward to continuing our collaboration in the future.

I would also like to thank Leonardo Basso, Jan Brueckner, José Luis Moraga, Jaap de Wit and Cees Withagen for being part of my thesis committee, for devoting time to the manuscript and for attending the defense. It is also a pleasure to have Robin Lindsey as a co-author of one of the papers that are part of this thesis; I hope we continue collaborating in the future. Leonardo Basso played an important role in my development as a researcher. His guidance during my year as a master student and during the time I was his research assistant has been crucial. Thanks Leo for your advice, for sharing your thoughts on much more than science, for encouraging me to take this step in my life and mostly for becoming my friend.

The Department of Spatial Economics is a very special place. I was lucky to spend my time as a PhD candidate with such a great group of people. I am indebted to the secretaries, but especially to Elfie and Jenny, for being welcoming, nice and friendly, for
making my life easier in so many occasion and for always checking how my family was
doing. You were always willing to help me and I appreciate that very much. I enjoyed
sharing office with people in the 4th and 9th floor over the years; special thanks to Jan,
Vincent, Ruben, Harry, Masha, Alexandros, Panayiotis, Martin, Ioannis and Masood. I
owe gratitude to the ones who kindly received my wife and me into their houses and
shared excellent food, drinks and time with us. Thanks Alexandros and Kelly for treating
us to Greek delicacies, Harry and Jeanine for Dutch traditional food, Masha for leading
an unforgettable supper in St. Petersburg, Achim for introducing me to Swabian cuisine
and to Islay whisky and Eva for sharing your mother’s lasagna with us. Thanks for these
and many other moments of joy. I also very much enjoyed the squash evening sessions at
Zuiver and being part of the VU-RE futsal team. Perhaps I enjoyed even more the beer
tasting events at the Department; thanks Harry for organizing this, for letting me be part
of it and thanks to the colleagues that joined and made them a remarkable experience. I
am also grateful to Harry for all the help provided in dealing with the Belastingdienst in
countless occasions.

I was lucky to make friends also outside the academic world in the last four years and I
am grateful for all the time we spent together. Among this group of people I would like to
especially thank David, Francesco and Sandra. Thanks David for the many coffee breaks
we had at the VU, for our weekly supper during the Uilenstede era, for making me an
Arsenal fan, for becoming a Colo-Colo supporter, for going with me to many Champions
League matches in different countries, for traveling with us to the Chilean Patagonia and for so much more. Thank you Fran and
Sandra for starting my latest birthday tradition, for those amazing boat rides along the
canals of Amsterdam in the *peque-peque*, for celebrating Queen’s day and King’s day with
us, for visiting Matías at the hospital, for all the dinners we had together and for being
the good friends you are.

Another source of happiness and motivation during these years was meeting old friends.
Thanks to those who were eager to meet up every time I was visiting Chile and made
those trips even better. Also thanks to those who visited us in Amsterdam and to those
who met us in our trips within Europe. I appreciate your company and I am very glad
that our friendship has survived the distance. For the sake of brevity I don’t mention you
all, but I cannot avoid showing my appreciation explicitly to Nino, Edú, Gabo, Emilio,
Juan, Sole, Thomas, Lalo, Fabi, Lautaro and Seba. Thanks Tía Sixti for providing us
with a home in Paris and for spending time with us.

I thank all my relatives and in-laws for their support and interest in what I was doing
these years and for hosting me during my visits to Chile. Accommodation, trips, dinners,
drinks, *asados*, football and much more I enjoyed with you. I especially want to mention
my grandmother Nena and my grandfather Hernán, who I lost in 2014 and who would
have been very proud of my achievement. I am lucky to have my sister Fernanda and my
brother Benjamín in my life. I am happy I could grow up with you and thank you for
being the way you are. I hope that Matías can have as much fun with you as I did, that
he can learn as much from you as I did and that we continue to be in each other’s lives.
I am deeply grateful to my parents Rebeca and Hugo. Without their support, lessons,
encouragement, concerns and love I would not be who I am. I learned a great deal from both of you. I hope I can continue making you proud and that you enjoy the new period of your lives as grandparents.

I dedicate this thesis to Daniela, the love of my life, and to Matías, my son and source of joy. Who would have believed, ten years ago, when Pata and I were dating, that we were going to be married and living in Amsterdam as proud parents of a wonderful child? Thank you for making the years that we have been together perfect, for leaving Chile (temporarily) behind, for being my wife and the mother of Matías. Thanks also for your approval, comfort and encouragement, for your constructive criticism, for listening to my presentations and for lifting my spirits. Thank you for being my partner in many adventures and for being there for me very single moment I needed you. Thanks for understanding my jokes and laughing at them, for your beautiful smile and for making me happy every day. I cannot imagine a life without you, I will not have a life without you.

Hugo

Amsterdam, February 2015
## Contents

**Preface**  

**Contents**

1 **Introduction**  
   1.1 Air transport  
   1.2 The economics of airport pricing  
   1.3 The industrial organization of transport markets  
   1.4 Outline of the thesis and preview of results

2 **Airlines’ strategic interactions and airport pricing in a dynamic bottleneck model of congestion**  
   2.1 Introduction  
   2.2 The model  
   2.3 A Stackelberg leader with a competitive fringe  
   2.4 Conclusions  

Appendix 2.A Glossary of notation  
Appendix 2.B Derivation of the equilibrium in the perfect competitive case  
Appendix 2.C Fringe’s response and leader’s first-order conditions

3 **On the existence and uniqueness of equilibrium in the bottleneck model with atomic users**  
   3.1 Introduction  
   3.2 The bottleneck model with homogeneous non-atomic users  
   3.3 Existence and non-existence of equilibrium with homogeneous atomic users  
   3.4 Nonuniqueness of equilibrium with heterogeneous atomic users  
   3.5 Conclusions  

Appendix 3.A Glossary of notation  
Appendix 3.B Proof of Lemma 3.2  
Appendix 3.C Proof of Lemma 3.4  
Appendix 3.D Proposition 3.2: Proof of Result 4 for mass departures  
Appendix 3.E Asymmetric equilibrium costs with $\gamma \leq \alpha$  
Appendix 3.F Heterogeneous desired arrival times

4 **Optimal pricing of flights and passengers at congested airports and the efficiency of atomistic charges**  
   4.1 Introduction  
   4.2 Airlines’ duopoly model
CONTENTS

4.3 Airport pricing and capacity investment ................................................. 93
4.4 Numerical analysis .................................................................................. 98
4.5 Conclusions .............................................................................................. 105
Appendix 4.A Glossary of notation ................................................................ 107
Appendix 4.B Reaction functions .................................................................. 107
Appendix 4.C Welfare maximizing airport first-order conditions .................. 109

5 Input third-degree price discrimination by congestible facilities ............. 111
  5.1 Introduction ............................................................................................... 112
  5.2 The model and the downstream markets .................................................. 115
  5.3 Private facility ......................................................................................... 117
  5.4 Public facility ........................................................................................... 125
  5.5 Comparison of the welfare effect under private and public ownership .. 129
  5.6 Robustness: downstream first-degree price discrimination .................... 132
  5.7 Conclusions .............................................................................................. 134
Appendix 5.A Glossary of notation ................................................................ 137
Appendix 5.B Calculations and proofs for Section 5.3 .................................... 137
Appendix 5.C Calculations and proofs for Section 5.4 .................................... 144
Appendix 5.D Calculations and proofs for Section 5.5 .................................... 148

6 Airline route structure competition and network policy ......................... 155
  6.1 Introduction ............................................................................................... 156
  6.2 The model ................................................................................................ 159
  6.3 Airlines equilibrium .................................................................................. 161
  6.4 Welfare analysis ....................................................................................... 170
  6.5 Sufficient instruments for social welfare maximization ......................... 176
  6.6 Conclusions .............................................................................................. 183
Appendix 6.A Glossary of notation ................................................................ 185
Appendix 6.B Functional forms and parameters of the main case ............... 185
Appendix 6.C Sensitivity analysis ................................................................... 186

7 Conclusions ................................................................................................. 193
  7.1 Summary of results and implications ....................................................... 194
  7.2 Suggestions for future work ..................................................................... 197

Bibliography .................................................................................................... 201

Samenvatting (summary in Dutch) ................................................................. 211
Chapter 1

Introduction
Chapter 1 Introduction

1.1 Air transport

Air transport has been at the core of the economy and of the daily life for some decades now. Nowadays, aviation is an essential part of the economic activity: firms, households, tourism and trade rely, to a substantial and growing extent, on air transportation. For example, since 1980 the number of passengers transported by air has increased more than fourfold and, in 2012, airlines transported over 3 billion passengers (ATAG, 2014). In the same year, 52% of international tourists traveled by air and 50 million tonnes of freight were carried in an aircraft. The total value of the goods carried by air represented 35% of all international trade and the aviation’s global economic impact is estimated at $2.4 trillion, which is 3.4% of the world gross domestic product (ATAG, 2014). Nevertheless, the growth of air transportation has not come without a cost. Air transport delays are an acute problem globally. For example, in the first half of 2007, 30% of the commercial flights in the US arrived more than 15 minutes late (Rupp, 2009). In 2005, over 20% of all intra-European flights departed more than 15 minutes late (Santos and Robin, 2010). Similar figures can also be found in other aviation markets. Ball et al. (2010) estimate that the cost of US air transportation delay in 2007 was $16.7 billions to passengers, $8.3 billions to air carriers, and that it reduced the 2007 US GDP by $4 billion. Not surprisingly, policies aimed at reducing airport congestion costs have gained increased attention from governments, policymakers and researchers.

The desirability of an intervention in this market or, at least, the economic rationale behind it, is partly based on the observation that delays are a consequence of the presence of negative externalities. In this market, the externality arises because there is a cost of traveling that is not borne by the passenger that travels, and a cost of scheduling a flight at a particular moment that is not borne by the airline that schedules the flight. An additional flight or passenger imposes additional travel time costs on other flights and passengers through interactions on the runway, in the air, in security screening areas, or gateways. This is normally referred as the congestion externality. Other negative externalities are also present in air transport markets: noise, air pollution and carbon emissions are main examples. The problem caused by congestion externalities has certainly been addressed before. Pigou (1920) and Knight (1924) already recognized its presence in road traffic, and since then we know how this market failure leads to a sub-optimal outcome. Pigou (1920) discussed how the principle of marginal-cost pricing applies to that case, where the free-market level of consumption (e.g. number of trips) is higher than socially optimal. This basic concept has been extended by many authors (e.g. Walters, 1961; Mohring and Harwitz, 1962; Vickrey, 1963), giving rise to an enormous road congestion pricing literature.

Solutions to the airport congestion problem have been intensively discussed in the literature. One option is to increase the capacity of congested airports. As Winston (2013) discusses in the context of US air transportation, although there is a large potential for welfare improvement through investments in airport runway capacity, especially when combined with efficiency pricing, the cost of constructing runways has become too large because it takes decades to meet regulations such as environmental impact standards. Runway capacity expansions bring benefits after a long period of time and at a relatively high cost. Another approach to deal with airport congestion is to control the total flight
volume through slot constraints. Under this scheme, a regulator fixes a number of landing and takeoffs that are allowed to take place in a specific time window, the slot window, and allocates the slots among airlines. This system is used in most European airports, many Australian airports and in a few US airports, among other places. It is often said that the efficiency of this approach relies heavily on the efficiency of the allocation of slots and it has been argued that by having a secondary market of slot sales and trade, the efficient outcome would arise (e.g. Brueckner, 2009). In fact, the current allocation system in Europe is in the process of being adapted towards having the possibility for secondary slot trade (European Commission, 2011). However, slot auctions and trade may lead to increased market power that could cause inefficient market foreclosure and entry barriers, as Verhoef (2010) discusses. In a recent study, Fukui (2014) finds evidence of a negative effect of slot trade between rival carriers on the number of competitors at a route level, which suggests the presence of anticompetitive effects. Even if these problems did not arise, recent contributions to the analysis have shown that slots may be inefficient. Daniel (2011) empirically finds that slot-constraints at Toronto airport are ineffective and that airport congestion pricing outperforms the slot constraint approach. Furthermore, Daniel (2014) shows theoretically that for a slot constraint policy to be effective, it requires many narrow slot windows, which makes the slot auctions and secondary markets difficult to implement.

A third alternative to address the increasing delays at airports is congestion pricing. This is the policy path that this thesis is concerned with. The objective of this thesis is to study socially optimal airport pricing policies in the presence of congestion externalities and in the presence of market power exertion by airlines. That is, to study pricing policies that are concerned with eliminating the excessive delays, recognizing that airlines are not price takers, but large agents that possess the ability to affect prices and delays. This thesis, therefore, models the consumers’ choice of traveling and the airlines’ behavior to study efficient airport pricing policies.

1.2 The economics of airport pricing

The current practice of levying takeoff and landing fees based on the weight of the aircraft that are constant throughout the day has been criticized since early contributions by Levine (1969) and Carlin and Park (1970). They were the first to contend that such practice is inefficient and that it leads to low load factors and excessive delays. Since then, several theoretical and empirical analyses have investigated efficient runway pricing policies and their potential benefits, and have argued in favor of its implementation. Nowadays, the general consensus is that the extensive congestion pricing literature developed for road traffic may not be directly applicable to airport congestion. The difference lies in that airlines usually control large shares of traffic at some airports and have market power, in contrast to road drivers who control a single vehicle. Airlines are large players in the aviation markets and, therefore, are expected to internalize the congestion imposed on themselves and react in a different way than road users when facing congestion charges. Early evidence that supports the presence of market power exertion by airlines is provided by Brander and Zhang (1990), who study empirically the airlines’ conduct in a set
of US routes. They infer whether carriers behave as price takers or if the pricing behavior implied by the data fits any of the other existing theories of firm behavior (e.g. cartel pricing). They conclude that the model of competition introduced by Cournot (1838) is consistent with the data. The conclusions are further extended by Oum et al. (1993). Based on these two studies, the assumption that airlines behave as Cournot oligopolists has become the standard in the analytical airport pricing literature.

The theoretical literature on airport pricing that explicitly models the airline market as an oligopoly where carriers have market power, as Basso and Zhang (2007b) point out, starts with the contribution by Brueckner (2002). As a matter of fact, Basso and Zhang (2008) show that a large share of the previous literature on airport pricing only applies to the case where airlines behave as price takers. Brueckner (2002) analyzes the effect of airline market power on airport congestion and the resulting socially optimal congestion charges. He shows that airlines behaving as Cournot oligopolists internalize the congestion imposed on themselves and that the efficient congestion toll should consider only the uninternalized fraction of congestion, which is the delays imposed on the other airlines. Pels and Verhoef (2004) extend the analysis by explicitly modeling the effect of the airlines’ market power on the ticket prices. They show that the socially optimal airport charge should account for two counteracting effects: the airlines’ incentives to contract output in order to increase profit –the traditional market power effect– and the lack of full congestion internalization, which leads to output levels that are higher than optimal. As a result, if the only regulatory instrument available is pricing, the efficient charge consists of the marginal congestion cost imposed on the other airlines and of a subsidy that neutralizes the market power markup.

These findings have important policy implications. First, the socially optimal airport charges will be typically highly differentiated, as an airline’s congestion charge should be proportional to the aggregate market share of the other airlines (Brueckner, 2002). This may undermine the implementation feasibility of the socially optimal pricing policy because airlines with large market shares should pay significantly less than small airlines, something that may be perceived as inequitable. Second, as Pels and Verhoef (2004) point out, the resulting efficient airport charge may be negative. If subsidies are unfeasible, which they normally are, the second-best policy would be to abstain from charging airlines. The pricing rule with these two counteracting components (market power subsidy and congestion toll) –a direct consequence of the Cournot assumption– has been shown to remain valid under several different settings. These include a network setting (Brueckner, 2005), when airport capacity investments are included (Zhang and Zhang, 2006), in presence of airport substitution (Basso and Zhang, 2007a) and when airlines offer differentiated products and schedule delay costs are considered (Basso, 2008). For a more detailed review of the airport pricing literature we refer to Zhang and Czerny (2012).

The result that airlines internalize the delay imposed on themselves also suggests that delays at highly concentrated airports should be at nearly efficient levels. However, this is in sharp contrast with the high levels of congestion observed in, for example, concentrated hub airports. This apparent contradiction gave rise to the so-called internalization debate. Daniel (1995) was the first to recognize that airlines may internalize delays and tested the hypothesis with data from Minneapolis-St.Paul (MSP) airport by comparing actual
and predicted traffic patterns of a stochastic dynamic model of congestion. Daniel and Harback (2008) extend the analysis to the 27 largest US airports and find that their specification tests largely reject the internalization hypothesis in most of the airports. Another branch of the literature uses a different empirical modeling approach and studies the relationship between delays and airport concentration. In this line, Brueckner (2002), Mayer and Sinai (2003) and Ater (2012) find that delays are lower at concentrated airports in the US and Santos and Robin (2010) in European airports, a result that supports the internalization hypothesis. Rupp (2009), on the other hand, using excess travel time as a measure of flight delays, finds that the results are reversed with US data.

This on-going debate, which determines whether congestion charges should be high as if airlines did not internalize congestion or they should be discounted proportionally to the airport share of flights of an airline, has been addressed from the theoretical and empirical sides. Morrison and Winston (2007) estimate the benefits of implementing both congestion pricing schemes at US airports. They find that most of the benefit from the socially optimal policy—which discounts the internalized congestion—can be achieved by charging carriers as if they did not internalize congestion. Therefore, fine-tuning the tolls according to (presumably) internalized congestion yields small benefits. However, these conclusions are drawn by assuming that airlines are price takers and therefore omits the argument, put forward by Pels and Verhoef (2004), that in concentrated markets the output could be below the optimal level. Brueckner and Van Dender (2008) and Daniel (2009) show theoretically that the outcome of the competition between a firm that acts as Stackelberg leader and a competitive fringe can approach the outcome of a model in which no airline internalizes congestion. This is likely to happen if the fringe and leader are perceived as perfect substitutes. This, however, cannot explain differences in findings between models and case-studies that consider simultaneous competition. Therefore, the question on the desirability and shape of a congestion pricing policy is still open.

This thesis aims to provide new insights into the airport pricing policy debate. The thesis is concerned with studying whether the feasibility problem of optimal pricing due to highly differentiated charges, the second-best optimality of a *laissez-faire* policy and the apparent inconsistency between theory and practice regarding congestion internalization are intrinsic to the air transport industry or they are consequence of the modeling strategy. It therefore aims at shedding light on questions such as: is there scope for implementing congestion pricing at concentrated airports where airlines are not price takers; and: is the result that delays are lower at more concentrated airports inconsistent with observed delay patterns? This thesis is also concerned with investigating broader pricing policies. It contributes to identifying when is efficient to implement a policy that bans airports from setting differentiated charges to airlines. It also analyzes the socially optimal pricing policy in a setting where the different pricing schemes may affect the airlines’ choice of how to serve the markets in a network, and whether an instrument directly targeted at regulating the route structure choice of airlines is needed to maximize social welfare.
1.3 The industrial organization of transport markets

The thesis also contributes to a broader strand of the economics literature by studying the industrial organization of transport markets. The field of industrial organization is mainly concerned with studying the structure and behavior of firms, and the functioning of markets and their efficiency (see Tirole (1988) for an overview of the field). It lays the basis for assessing the scope for government intervention and for studying policies aimed at improving the efficiency of markets. In that sense, when studying the functioning of transport markets, this thesis contributes to the field of industrial organization. Within the industrial organization of transport markets, aviation—the subject of this thesis—arguably the most studied topic, yet the literature is broader. Important contributions exist in the areas of rail transportation, provision of private roads and public transportation, and accident insurance among others (see e.g. Small and Verhoef (2007) for an overview of the contributions in these markets).

Usually, and especially in the case of air transportation, as discussed above, there is imperfect competition between the firms that offer the final good. Consequently, the thesis relies on oligopoly theories to model the transport markets and derive policy insights. In particular, it uses and extends the contributions by Cournot (1838), Bertrand (1883) and von Stackelberg (1934) to represent oligopolistic markets. The thesis also builds upon the models of Dixit (1979) and Singh and Vives (1984) to study product differentiation using a “non-address” approach, where firms produce differentiated goods and a consumer that is representative of the population has preferences over the goods and a taste for variety. It also incorporates into the models of oligopolistic competition features that are typical of transport markets, such as negative consumption externalities, which play a crucial role in the analysis. For example, as the thesis shows, different combinations of assumptions about the firms’ conduct and the congestion technology may yield fundamentally different policy insights.

The vertical nature of the aviation markets, where airports are input providers and airlines are downstream firms, makes the analysis closely related to the literature on vertical externalities. A profit maximizing airport does not take into account the incremental profit of the airlines when making its decisions, which results in the so-called double marginalization problem (Spengler, 1950). When the input provider, instead of maximizing profit, maximizes domestic welfare because, for example, it is owned by a national authority and serves international markets, another type of vertical externality arises. This national airport takes into account the incremental profit of domestic firms, but not that of foreign firms, and does not take into account the foreign consumer surplus. On the other end of the ownership spectrum, a welfare-maximizing upstream provider may achieve the efficient outcome using a linear price if the downstream firms view the price as parametric (as opposed to believing that they can influence the upstream price with their actions). The latter case is useful to obtain insights into how the airport charges should be set, and the other cases, more realistic indeed, can be useful for analyzing how to implement those welfare-maximizing prices.

Although the thesis often explicitly refers to airports as input providers and airlines as downstream firms, the results and conclusions can readily be extended to other (transport) markets. As long as these markets are also well described by a vertical structure with
negative consumption externalities, the theoretical insights obtained in this thesis are also applicable to those markets. Examples of such markets include rail transportation, sea transportation, and, in some cases, telecommunication markets.

The thesis also extends the models of industrial organization in networks where product complementarity and substitutability are crucial in shaping the market outcomes (e.g. Economides and Salop, 1992). Transport markets have an intrinsic network aspect: a trip, by definition, has an origin and a destination. In these markets, firms can provide substitute products as well as complement products (e.g. airline alliances), but also the different providers of the transport infrastructure can offer substitutes or complements (e.g. roads, train stations, airports). By explicitly modeling these aspects, the thesis contributes to understanding the functioning of transport network markets.

In addition, the thesis contributes to the literature on price discrimination. There is a vast literature that studies the effects that allowing price discrimination has on prices, quantities, and welfare that dates back, at least, to the contribution by Robinson (1933). However, the analysis in presence of negative consumption externalities, which is a typical feature of transport markets, is relatively scarce. Studies that do analyze this focus on price discrimination in final markets (e.g. Adachi, 2005; Czerny and Zhang, 2015). The thesis contributes to this strand of the literature by studying input price discrimination by private (profit maximizing) and public (domestic-welfare maximizing) suppliers in the presence of negative consumption externalities.

1.4 Outline of the thesis and preview of results

This thesis proposes different theoretical models to answer various research questions. Chapters 2 and 3 investigate the airlines’ scheduling behavior in a dynamic model of congestion when they control large shares of traffic at a facility and analyze efficient runway pricing in this context. The chapters look at different market structures and at the properties of the different equilibria, cast light on the sources of inefficiency of the airlines’ scheduling decisions and how to deal with these by means of setting (time-varying) charges. A main objective is to study whether the results from static models of congestion are structurally different from those obtained when the congestion technology is assumed to be dynamic. For this purpose, both chapters adopt a short-run view, focus on within-the-peak dynamics, treat aircraft size as given, and neglect network effects. The thesis then moves on to structural longer run models, where static (flow) congestion is assumed to apply. Chapter 4 studies airport pricing policies under different assumptions on the airlines’ market conduct and, thus, departs from the traditional theoretical works based on Cournot models. By making aircraft size endogenous and comparing the outcomes of different airline behavioral assumptions, the chapter contributes to the internalization debate and to reassessing the (second-best) optimality of a laissez-faire policy. Next, chapter 5 sheds light on a broader policy question: should price discrimination by airports be banned? That chapter, therefore, focuses on studying the effects of allowing different types of airports to set differentiated prices, as opposed to describing the socially optimal prices that should be charged. Finally, Chapter 6 deals with airport optimal pricing policies in a long-run model where airlines adapt the way they serve a network according
to the different airport charges. The chapter analyzes how the conclusions from shorter-run models apply in a network framework. The outline of the technical chapters of the thesis, and the main characteristics, are summarized in Figure 1.1.

Chapter 2 analyzes efficient pricing at a congested airport dominated by a single firm. Unlike much of the previous literature, the chapter combines a dynamic bottleneck model of congestion and a vertical structure model that explicitly considers airlines and passengers. The analysis confirms that a Stackelberg leader interacting with a competitive fringe partially internalizes congestion and shows that there are various toll regimes that induce the welfare maximizing outcome, widening the set of choices for regulators. In particular, charging the congestion toll that would apply for fully competitive carriers and that ignores any internalization, to both the leader and the fringe, yields the first-best outcome.
In Chapter 3, the analysis of the previous chapter is extended by considering simultaneous fleet scheduling by firms. This chapter investigates the existence, uniqueness and properties of equilibrium in the Vickrey bottleneck model when each firm controls a sizable fraction of total traffic. Firms simultaneously choose departure schedules for their vehicle fleets and each firm internalizes the congestion cost that each of its vehicles imposes on other vehicles in its fleet. The chapter establishes three results. First, a pure strategy Nash equilibrium (PSNE) may not exist. Second, if a PSNE does exist, identical firms may incur appreciably different equilibrium costs. Finally, a multiplicity of PSNE can exist in which no queuing occurs but departures begin earlier or later than in the social optimum. The order in which firms depart can be suboptimal as well. Nevertheless, the chapter shows that by internalizing self-imposed congestion costs, individual firms can realize much, and possibly all, of the potential cost savings from either centralized traffic control or time-varying congestion tolls.

Chapter 4 investigates and compares airport pricing policies under various types of downstream competition, considering both per-passenger and per-flight charges at congested airports. The chapter shows that an airport requires both pricing instruments to achieve the first-best outcome and distinguishes their role by showing that congestion externalities need to be addressed through per-flight tolls whereas the inefficiency caused by the airlines’ market power exertion must be corrected with per-passenger subsidies. The chapter also shows that Bertrand competition with differentiated products, a type of behavior recently pointed out by the empirical literature as pertinent, has policy implications that diverge from analyses that assume Cournot competition. The welfare gains and congestion reductions of congestion pricing under a Bertrand oligopoly would be higher than what has been advanced before under a Cournot oligopoly; the degree of self-financing of airport infrastructure under optimal pricing would be increased and may approach exact self-financing; and the implied differentiation of charges between (asymmetric) airlines would be significantly smaller, presumably enhancing the political feasibility of welfare maximizing congestion pricing, as the potential distributional concerns would be decreased. Finally, the chapter ends with a numerical analysis of second-best policies and finds that pricing airlines as if they did not internalize congestion may offer a relatively attractive alternative to first-best congestion pricing.

Chapter 5 studies third-degree price discrimination by transport facilities, such as airports, which sell access to the infrastructure as a necessary input for downstream production. These facilities are prone to congestion –which makes downstream markets interrelated– and their ownership structure is diverse, varying from public (domestic welfare maximizing) to private (profit maximizing). The analysis shows that allowing input price discrimination by a private supplier can increase aggregate output and increase welfare in a setting where, in absence of congestion, output does not change and welfare is reduced when price discrimination is allowed. Therefore, the presence of negative consumption externalities enlarges the extent to which input price discrimination is desirable. The chapter also describes the conditions under which banning input price discrimination is efficient for both types of ownership form of the facility, and argues that there is a limited scope for this to happen. This suggests that the current practice of enforcing a broad ban on input price discrimination that covers congestible facilities with different ownership forms may have to be revised.
Chapter 1 Introduction

Chapter 6 analyzes the behavior of airlines in terms of route structure choice using a differentiated duopoly model that accounts for congestion externalities, passenger benefits from increased frequency, passenger connecting costs and airline endogenous hub location. The chapter also examines the route structure configuration that maximizes welfare, and studies whether it can or will arise as an equilibrium when a regulator implements socially optimal airport pricing, but does not regulate directly the route structure choice. The analysis finds that a separate instrument directly aimed at regulating route structure choice may be needed to maximize social welfare, in addition to per-passenger and per-flight tolls designed to correct output inefficiencies given the network structure. This holds true when the regulator is constrained to set non-negative tolls, but also for unconstrained tolling. Finally, the chapter also studies the relative efficiency of airport pricing when the optimal route structure configuration cannot be decentralized by tolling.

Chapter 7 concludes by summarizing the main findings, and offering policy advices that can be drawn from this thesis as well as guidelines for future research.
Chapter 2

Airlines’ strategic interactions and airport pricing in a dynamic bottleneck model of congestion
2.1 Introduction

As congestion at major airports worldwide continues to increase and traffic approaches existing capacities, implementing policies aimed at reducing delays effectively is becoming essential. For example, in the first half of 2007, 30 percent of commercial flights in U.S. arrived more than 15 minutes late, and similar figures hold for European airports (Rupp, 2009; Santos and Robin, 2010). Policies to solve the congestion problem have been extensively discussed during the last decades. One alternative is capacity enlargements, but these have the drawback of bringing benefits only after a long period of time, and at a relatively high cost (see Jorge and de Rus (2004) for a cost-benefit analysis). Another option is congestion pricing, perhaps the most discussed policy in the academic economics literature, often heavily inspired by the road pricing literature. However, governments, regulators and airports have not followed this path. The current practice at many airports is to levy weight-based landing fees, a rule that has been criticized since early contributions by Levine (1969) and Carlin and Park (1970), who were the first to argue that these charges provide wrong incentives and lead to inefficiencies. Despite of four decades of theoretical and empirical contributions calling for implementation of efficient landing and takeoff charges based on economic principles, airport pricing schemes have been kept remarkably unchanged. But, as delays are reaching critical levels and other negative externalities, such as pollution and noise, are becoming more important, congestion pricing is likely to turn into a serious option for governments and regulators. This policy may be specially appealing because landing fees are already in place, and only changes are needed in the way that they are charged. Moreover, in some countries, such as the U.S., landing fees are allowed to vary by time of the day, a fundamental feature of an efficient congestion pricing scheme.

It is now widely agreed that the vast literature on road congestion pricing may not be directly applicable to airports, because airlines are non-atomistic players, in contrast to road drivers. Carriers have market power and have non-negligible shares of the overall traffic and, as a consequence, they can be expected to internalize the congestion imposed on themselves. Daniel (1995) was the first to recognize this, and Brueckner (2002) and Pels and Verhoef (2004) analyzed the problem assessing the internalization of congestion with theoretical models. Subsequent works by Brueckner (2005), Zhang and Zhang (2006), and Basso and Zhang (2007a) extend the analysis. The main conclusion regarding congestion pricing, based on static models of congestion, is that carriers competing in a Cournot-Nash fashion internalize self-imposed congestion and, therefore, should be charged for the fraction of congestion that they impose on others. This leads to a congestion charge

---

1This chapter is based on joint work with Erik Verhoef and Vincent van den Berg. An earlier version of this chapter has been published in Journal of Urban Economics (Silva et al., 2014c). I am indebted to Achim Czerny, Tomás González, Sergejs Gubins and Eric Pels for helpful comments and suggestions. The main variables and parameters that are used in this chapter are summarized in Appendix 2.A.

2Quantity-based approaches to congestion management are also being discussed as an alternative. See Brueckner (2009), Basso and Zhang (2010) and Verhoef (2010) for analyses on slot sales and slot trading.

3Congestion pricing can be a second-best solution for environmental externalities. See, for example, Carlsson (2003) for an analysis of airport pricing with congestion and emissions, and Brueckner and Girvin (2008) for an investigation of airport noise regulation.
2.1 Introduction

that depends on the rivals’ market share at the congested airport, and, therefore, may be perceived as inequitable, as dominant airlines should face lower charges than small carriers.

The contribution of this chapter is to provide clear-cut insights into and understanding of airlines’ strategic interactions and airport congestion pricing in a model of dynamic congestion. We recognize the vertical nature of aviation markets, thus explicitly including the role of the airport’s tolls on the airlines’ behavior, and incorporating that airlines compete taking these into account, while facing the passengers’ demand for trips. We use the deterministic bottleneck model of congestion developed by Vickrey (1969) and Arnott et al. (1990, 1993). This allows for an analysis that balances analytical tractability and the inclusion of behavioral decisions that we believe are essential: airlines endogenously adjust departure or arrival rates, trading off queuing delays and schedule delays, and passengers dislike queuing and schedule delays in a different manner than airlines (i.e. at different shadow prices). By combining these two modeling features, we have a structural model of dynamic congestion that allows for an analysis of the firms’ inefficiency in terms of the number of flights as well as the scheduling, and, as a consequence, allows for a derivation of the optimal policy that deals with both. We focus on sequential competition between a Stackelberg leader and a competitive fringe. The model set-up is consistent with the empirical findings of Daniel and Harback (2008), who show that observed traffic patterns at most of the major U.S. airports are consistent with the dynamic bottleneck model of congestion, and that most of the U.S. hub airports seem best described by competition between a Stackelberg leader and a competitive fringe.

Our main result is that, while the (untolled) equilibrium is fully consistent with what previous literature with static congestion suggests, first-best congestion pricing is not. In particular, when a Stackelberg leader faces a competitive fringe, the equilibrium is fully consistent with static models in that the fringe does not internalize any congestion, and in that the leader’s ability to exert market power and to internalize self-imposed congestion depends critically on the assumed substitution pattern (just as in Brueckner and Van Dender (2008)). On the other hand, we find that the first-best optimum can be decentralized with a pricing policy that consists of a market power subsidy for the leader, that is indeed a function of the assumed substitution pattern, and a congestion toll for both agents that is independent of whether internalization occurs in the untolled setting. We show that charging the congestion toll that is derived for the fully atomistic carriers to both leader and fringe always yields the first-best outcome. This is because the subsidy deals with the leader’s overpricing due to market power, and the time-varying congestion toll eliminates queuing and provides the right incentives to take into account the delays imposed on the rival airlines. We further show that there are various alternative toll regimes that also attain the first-best, dealing with the congestion inefficiency in yet different ways, while still correcting for the market power exertion. Again, the congestion component of all toll regimes is independent of the degree of internalization by the leader in the unregulated equilibrium.

The results of this chapter suggest that optimal congestion pricing may have a more significant role on airports than what has been suggested in the literature before. The congestion pricing scheme that is obtained for fully atomistic carriers induces the first-best outcome, and results in a revenue for the airport that restores the well known self-financing
result for congested facilities: the ratio between first-best capacity investment costs and total revenue from congestion pricing equals the degree of economies of scale in capacity provision (Mohring and Harwitz, 1962). In addition, our results suggest that the political feasibility of optimal congestion pricing would be enhanced, as the (first-best) atomistic congestion charges do not vary across airlines and therefore are less likely to be perceived as inequitable. Finally, the fact that there are several tolling regimes that yield the social welfare maximizing outcome widens the set of choices for regulators.

Our analysis contributes to the policy analysis on congested airports and extends previous literature that considers dynamic congestion at airports. Works such as Daniel (1995, 2001) and Daniel and Harback (2008, 2009) focus on cost minimization of scheduling flights, hence ignoring the passengers’ role in the problem, or at least treating that role only implicitly. Moreover, most of these papers aim at testing whether the observed patterns of arrivals and departures of flights support the internalization hypothesis. Daniel (2009) analytically studies the conditions under which dominant airlines internalize self-imposed congestion with a deterministic bottleneck model, focusing on Stackelberg-fringe competition, but omits the passengers in the model, hence ignoring the fact that airlines use the airport as an input to sell an output in a downstream market. By combining the bottleneck congestion model with the explicit consideration of two groups of agents (airlines and passengers) in a theoretical model, we are able to study key elements that were not present in previous exercises with dynamic congestion. These include an analysis on how airlines set the ticket price according to the time of departure, a derivation of an explicit relation between the internalization of congestion and the assumed passengers’ demand substitution pattern between airlines, and a clear comparison between the results derived in models of static congestion and the results obtained with dynamic congestion. We are also able to study the implications, for the optimal pricing policy, of the strategic interaction between the leader and the fringe, finding that there is a set of various pricing schemes that maximize social welfare, as opposed to a single optimal congestion toll. Finally, our analysis complements the findings of Brueckner and Van Dender (2008) and Silva and Verhoef (2013) who show that congestion charges can be optimally close to the atomistic charges depending on the assumptions on the prevailing market structure.

Our results have to be qualified according to our assumptions. Naturally, the dynamic bottleneck model is not directly applicable when queuing is not necessary or helpful for airlines in order to obtain a certain arrival time, as in fully slot-constrained airports. This is because the airport’s regulator directly controls the timing through slot allocations. For this case, more common in European airports, an analysis of slot sales and slot trading is more pertinent (see Brueckner 2009). We also assume that airlines and passengers share a most desired time of arrival or departure, and that airlines are homogeneous in

---

4 We also show how the market-power exertion has to be corrected, finding insights that are consistent with those in the previous literature, and that this overturns the self-financing result if market-specific subsidies are drawn from the airport budget.

5 Daniel (2009) recognizes that the dynamic atomistic toll charged to all airlines induces the welfare maximizing output in his scheduling model, but he does not analyze the leader’s response to the fringe behavior when facing the toll, and therefore does not find alternative schemes. He also omits the passengers’ role in the analysis, and our behavioral model seems to match his set of assumptions only when leader and fringe serve independent markets whose demands are related only through congestion.
2.2 The model

values of time. The model can be straightforwardly extended in these directions following
the road pricing literature.6 Lastly, we use the deterministic version of the bottleneck
model for analytical simplicity. A stochastic version that does not require attempted
inflows at or above capacity to yield queues would be more realistic. However, as the
trade off between expected queuing and expected schedule delays will be driving airlines’
interactions, general results may not change significantly, while detailed results such as
equilibrium delays, traffic rates and queue lengths will change.

The chapter is organized as follows. Section 2.2 introduces the model and the assump-
tions that are necessary for the analysis. We illustrate the main features of the model by
characterizing the untolled equilibrium and deriving first-best and time-invariant second-
best tolls for perfectly competitive airlines. We then study a monopoly carrier in the
market. Section 2.3 extends the analysis to competition where a Stackelberg leader faces
a group of competitive carriers, focusing on the untolled equilibrium and on first-best
tolling. We study the case of imperfectly elastic demand and imperfectly substitutable
airlines, and also look at the special cases of perfect substitution, independent markets
and perfectly elastic demand. Finally, Section 2.4 concludes.

2.2 The model

2.2.1 The basics and perfect competition

This section describes our model of dynamic airport congestion and shows how it incor-
porates the main features of airlines and passengers behavior, by looking at the case of
perfect competition. Subsequent sections look at the extension to different market struc-
tures. We base our analysis in the work of Vickrey (1969) and Arnott et al. (1990, 1993),
extending their dynamic congestion modeling for atomistic users (road users), to the case
where congestion occurs in a facility used by carriers with market power (airlines), who
sell their output (trips) in a downstream market of passengers. We focus our analysis
on arrivals, thus the bottleneck is the airport’s runway and queuing takes place in the
air, before landing. The analytical results would apply for departures as well, and can
in principle be extended to a network setting with multiple airports and delays in both
arrivals and departures.7

This model considers “pure” bottleneck congestion behind a bottleneck of finite capacity,
implying that in absence of a queue and, as long as the arrival rate of flights at the
bottleneck is below its capacity, there are no travel delays. Under other conditions, the
queuing delay experienced by a flight and its passengers depends on the length of the
queue at the moment of joining it. We assume that free-flow travel time is zero,8 so that a

6The original model by Vickrey (1969) analyses heterogeneity in desired arrival time. For heterogeneity
in values of time see e.g. Vickrey (1973), Arnott et al. (1994) and Van den Berg and Verhoef (2011).
7Note that our bottleneck model is relevant when the airport’s operational conditions for arrivals (or
departures) follow the first-in first-out (FIFO) discipline, and is not directly applicable when the
airport is managed with slots, because the airport’s regulator directly controls the timing through
slot allocations.
8In a single origin-destination pair, we can assume zero free-flow travel time without loss of generality,
but this is generally different with multiple origin-destination pairs.
flight that departs from the origin at \( t_d \), arrives at the bottleneck at the same time. As a consequence, in absence of queuing, the time of arrival at the destination (landing), \( t \), also matches the time of departure from the origin. When there is queuing, the arrival time, \( t \), is the time of departure (\( t_d \)) plus the queuing delay \( T(t) \), and the length of the queue, \( Q(t) \), grows or shrinks at a rate \( \dot{Q} = r_d - K \), where \( r_d \) is the aggregate departure rate and \( K \) the capacity of the bottleneck. The travel delay of a flight arriving at destination at \( t \), is the length of the queue at the moment of joining it, divided by the bottleneck’s capacity:

\[
T(t) = \frac{Q(t_d)}{K} \quad \text{with} \quad t_d = t - T(t) \tag{2.1}
\]

Note that this definition, due to the perfect information assumption, allows us to write time costs as a function of the arrival time at the destination, instead of the departure time from the origin.

We follow Small’s (1982) model of scheduling behavior for both passengers and airlines, so that their time costs are the sum of travel delay cost and schedule delay cost. Passengers, in a nutshell, experience travel delays in the form of queuing delays to land, and have a preferred arrival time \( t^* \) from which any deviation (early or late) induces a schedule delay. The passengers’ schedule delay cost, that arises from the difference between desired and actual arrival (or departure) time, was introduced in the context of aviation by Douglas and Miller (1974) and estimated by Morrison and Winston (1989) as part of a passengers’ discrete choice model of airline.\(^9\) A natural interpretation for this cost is that people, everything else constant, want to arrive at their destination at a certain moment, that can be, for instance, the start of the working day in order to make the most out of it.\(^\text{10}\) The schedule delay costs for airlines is a less studied matter. However, the scheduling of crew and coordination of arrivals and departures (specially in hub-and-spoke networks), are possible interpretations for including early and late schedule delay costs for airlines. In addition, as we show in Appendix 2.B, our analysis and results hold in absence of airline schedule delay costs. Phrased differently, an airline’s own schedule delay cost enters its maximization problem in the same way as its passengers’ schedule delay costs do, as the latter imply a decreased willingness to pay a ticket fare.

Let \( g \) be a sub-index that denotes agent-type (\( p \) for passengers and \( a \) for airlines), \( \alpha_g \) the value of travel time for agent type \( g \), \( \beta_g \) the value of early schedule delay, and \( \gamma_g \) the value of late schedule delay. Then, the generalized cost of arriving at \( t \), for an agent type \( g \), \( C_g(t) \), can be written as:

\[
C_g(t) = \alpha_g \cdot T(t) + \begin{cases} 
\beta_g \cdot (t^* - t) & \text{if } t < t^* \\
\gamma_g \cdot (t - t^*) & \text{if } t \geq t^*
\end{cases} \tag{2.2}
\]

\(^9\)Using a reduced-form for the schedule delay cost in models of static congestion is common in the aviation literature (e.g., Oum et al., 1995; Brueckner, 2004).

\(^\text{10}\)For example, this directly applies to business travel. It can be argued that for leisure passengers this also hold as well, as, everything else constant, they prefer to arrive at a certain time during the day. Another possible interpretation is related to transfers at hub airports. Although our model does not consider network effects, one can think of passengers using the flight for a transfer and, in that case, \( t^* \) would represent their most preferred moment to arrive at the hub airport (the time that makes the transfer possible without experiencing undesired waiting).
2.2 The model

The airline’s generalized cost differs from the user’s generalized cost only in the values of time, which reflects our assumption that airlines share the desired arrival time $t^*$ with the passengers.\(^{11}\)

Having described the congestion modeling, we can turn to the passengers’ demand specification, and the airlines’ costs and profit. We assume, for the perfectly competitive case, that passengers perceive airlines as perfect substitutes, and that the demand for an airline follows a linear inverse demand function:

$$d \left( \sum_i q_i \right) = A - B \cdot \sum_i q_i \quad (2.3)$$

which gives the marginal willingness to pay for traveling; $q_i$ is the number of passengers traveling with airline $i$; $A$ represents the maximum reservation price, and $B$ is the demand sensitivity parameter. We use the linear specification for analytical simplicity, but our results do not depend crucially on this.

The full price $p_i$ for a passenger traveling with airline $i$ is the sum of the fare ($\rho_i$) and the generalized cost experienced by the passenger. As we consider dynamic congestion, the various components of the generalized cost are generally not constant over time (see Eq. (2.2)). The condition for an equilibrium, where all flights are used by passengers and where passengers are indifferent between all the flights, is given by:

$$\rho_i(t) + C_p(t) = A - B \cdot \sum_i q_i \quad (2.4)$$

which is simply the full price of taking any airline $i$’s flight, that arrives at destination at time $t$, equals marginal willingness to pay. Recall that the generalized cost experienced by the passenger does not depend on the identity of the airline, but only on the time of arrival. The equilibrium condition in Eq. (2.4) implies that airlines charge different fares for flights scheduled at different times, except for flights whose users experience the same generalized cost. Forbes (2008) provides empirical evidence that airlines indeed charge lower fares when they face higher delays.

As usual in the airport pricing literature, we assume that the product of the load factor and the seat capacity is constant, so that the number of passengers per flight is given. The airlines’ costs consist of a time-invariant operating cost per flight $c_1$, a time-invariant operating cost per passenger $c_2$, and the time-variant cost $C_a(t)$ described in Eq. (2.2). Denoting the constant product between seat capacity and load factor as $s$, time-invariant costs can be expressed as a constant cost per flight $c = c_1 + s \cdot c_2$.\(^{12}\) With the cost

\(^{11}\)Although the preferred arrival time for airlines may be endogenous, following from desired arrival times for passengers, the analysis of this issue is beyond the scope of this chapter. With endogenous $t^*$, it can be expected that the airlines’ preference is significantly affected by the passengers’ preferred arrival time and will be close in practice. For example, in hub-and-spoke networks, airlines coordinate arrivals and departures to facilitate passenger connections. Cost advantages because of high passenger density may also drive airlines to adopt the passengers’ preferred arrival time.

\(^{12}\)Because of the fixed-proportions assumption, constant costs per passenger and per flight have the same effect, and can be aggregated. The same occurs when airlines are charged a landing fee; it does not matter if it is a per-passenger fee or a per-flight fee.
structure defined, we can analyze the equilibrium in the airline market and then study the regulator’s problem. This section looks at the perfectly competitive case, to illustrate the main features of the model.

In the case of imperfect competition, airlines would have as decision variables the number of flights (or prices), and the departure time of each flight. In order to analyze the perfectly competitive case, we assume that there is a continuum of small competitive airlines that can enter the market by scheduling a single flight at any time. Therefore, each competitive airline’s decision variable is the time of arrival \( t \), and the aggregate number of flights will be given by the zero-profit condition. The profit of an airline, that schedules its only flight to arrive at \( t \), is revenues minus costs:

\[
\pi(t) = s \cdot \rho_i(t) - C_a(t) - c - \tau(t)
\]

where \( \tau(t) \) is the time-variant per-flight toll (in this case, landing fee) that the regulator might charge to airlines. Denoting \( f \) as the aggregate number of flights, the total number of passengers is \( s \cdot f \), and using the interior equilibrium condition in Eq. (2.4), airline profit is:

\[
\pi(t) = s \cdot [A - B \cdot sf - C_p(t)] - C_a(t) - c - \tau(t)
\]

where the term between square brackets is the fare. Using Eq. (2.2) and defining \( \bar{\alpha} = s \cdot \alpha_p + \alpha_a \), \( \bar{\beta} = s \cdot \beta_p + \beta_a \) and \( \bar{\gamma} = s \cdot \gamma_p + \gamma_a \), the profit of an airline whose flight arrives at time \( t \) can be simplified as:

\[
\pi(t) = s \cdot [A - B \cdot sf] - c - \tau(t) - \bar{\alpha} \cdot T(t) - \begin{cases} 
\bar{\beta} \cdot (t^* - t) & \text{if } t < t^* \\
\bar{\gamma} \cdot (t - t^*) & \text{if } t \geq t^*
\end{cases}
\]

This reduced form shows that airlines take into account the generalized cost of its own passengers, because the lower the passengers’ generalized cost is, the higher the fare can be (see Eq. (2.4)), on a dollar-by-dollar basis. Therefore, we can interpret the airline’s problem as if they incur a generalized cost per flight, that is the sum of its own generalized cost, \( C_a(t) \), and the generalized cost of all the passengers on its flight, \( s \cdot C_p(t) \).

The dynamic equilibrium is such that an airline cannot increase its profit by changing the schedule of its single flight, for a given scheduling behavior of the other airlines. By looking at Eq. (2.7), this can only be achieved when every airline, i.e. every flight, faces the same sum of toll and generalized cost per flight (the travel and schedule delay cost terms on the right-hand side of Eq. (2.7)), because all other terms are time-invariant. This generalized cost per flight from the airlines’ perspective, is similar to the generalized costs typically found in the bottleneck road pricing literature for individual drivers (e.g., Arnott et al., 1990, 1993). A difference is that the values of time considered by the airline, for a single flight, are its own values of time plus the summed passengers’ values of time in that flight. But, through the use of the composite shadow prices \( \bar{\alpha}, \bar{\beta}, \) and \( \bar{\gamma} \), this

---

13 Choosing the number of flights is equivalent to setting the number of passengers (quantity) because the fixed-proportion assumption implies \( q_i = f_i \cdot s \).

14 An alternative interpretation of the perfectly competitive case is that airlines are not necessarily small, but they view price and congestion level as parametric. This would imply that their fare and time of departure is given by the zero-profit condition and they choose volumes.
2.2 The model

difference disappears from the formal model. This enables us to describe the equilibrium in schedules following the road pricing literature, and keep the discussion concise.

We first characterize the untolled equilibrium. The equilibrium condition then is that the generalized cost per flight \((C_a(t) + s \cdot C_p(t))\) must be constant over time during the period of operation; otherwise an airline would have an incentive to reschedule its flight and increase its profit. As shown in the road pricing literature, there is a unique aggregate queuing pattern that satisfies this equilibrium property, and this pattern defines the (equilibrium) scheduling behavior of the competitive airlines (see Appendix 2.B for the calculations and derivation of this result). Denote \(t_s\) as the (endogenous) first moment of operation, i.e. the time where the first flight arrives at destination, and \(t_e\) as the (endogenous) end of the operation period. The first flight departs at \(t_s\) and arrives at the same time, as there is no queue, incurring only an early schedule delay cost. The same holds for the last flight, at \(t_e\), incurring only a late schedule delay cost (if the last flight incurred queuing, its costs could be reduced by departing later and still arriving at the same moment). Arrivals are continuous in this model, and as a consequence, the duration of the peak period has to be \(f/K\), the total number of flights divided by the capacity of the bottleneck. From \(t_s\) onward, the queue evolves, growing up to a maximum level (just when a flight arrives at \(t^*\)) and then decreasing until it dissipates completely at \(t_e\), in the unique way that makes the generalized cost per flight constant over time. The resulting constant generalized cost per flight can be found by determining the equilibrium timing of the peak of duration \(f/K\), such that the schedule delay costs are the same for the first and last flight. This gives two conditions \((\bar{\beta} \cdot (t^* - t_s) = \bar{\gamma} \cdot (t_e - t^*) \land t_e - t_s = f/k)\) that are sufficient to determine the equilibrium generalized cost:

\[
C_a(t) + s \cdot C_p(t) = \frac{\bar{\delta} \cdot f}{K} \quad \forall \quad t \in [t_s, t_e] \tag{2.8}
\]

where \(\bar{\delta} = (\bar{\beta} \cdot \bar{\gamma})/(\bar{\beta} + \bar{\gamma})\). The equilibrium departure rates can be derived from equating the time derivative of Eq. (2.7) to zero. Note that the aggregate scheduling pattern is unique, but an individual airline’s scheduling is undefined due to the perfect competitive assumption. This yields an equilibrium profit (superscript \(e\)) for any airline of:

\[
\pi^e = s [A - B \cdot sf] - c - \frac{\bar{\delta} \cdot f}{K} \tag{2.9}
\]

Recall that airlines are indifferent between any arrival time \(t\) between \(t_s\) and \(t_e\), and passengers are indifferent between any flight, because the full price of all flights is constant, equal to \(A - B \cdot sf\), and given by:

\[
p_i = \rho_i(t) + C_p(t) = A - B \cdot sf = \frac{1}{s} \left( c + \frac{\bar{\delta} \cdot f}{K} \right) \tag{2.10}
\]

where the last equality comes from the zero-profit condition of the perfectly competitive case \((\pi^e = 0)\). The passengers’ full price in the no-toll equilibrium equals the airlines’ constant operating cost per passenger \((c/s)\) plus the generalized cost per flight divided

\[15\]See Appendix 2.B for a derivation of this result, and Arnott et al. (1990, 1993) for a detailed discussion.
by the number of passengers. The total generalized costs (or travel delay plus schedule delay costs) are the generalized costs per flight times the number of flights, $\delta \cdot f^2/K$, as in the road case. In Appendix 2.B we extend the analysis by looking how the equilibrium fare varies over time.

Figure 2.1 illustrates the no-toll equilibrium for the competitive case. The equilibrium is represented by the constant generalized cost per flight (from Eq. (2.8)), and the depiction of $s[A - B \cdot sf]$ satisfying Eq. (2.10). The only conditions on the values of time that are needed for this equilibrium to exist are that $\alpha > \beta > 0$ and $\gamma > 0$. As these values of time are made up of a combination of the passengers’ and the airline’s values of time, the interpretation for the condition is not immediately straightforward. In the case of passengers’ values of time, empirical evidence indicates that the conditions are satisfied, i.e. that the value of travel time is higher than the value of schedule delay early ($\alpha_p > \beta_p$), and that the value of schedule delay late is above zero ($\gamma_p > 0$) (see Morrison and Winston, 1989; Lijesen, 2006). In the case of airlines, to the best of our knowledge, there is no empirical evidence for the values of schedule delay. However, given that the passengers relation is intuitive and has empirical support, the only additional assumptions that we need on the airlines’ values of time are that $\alpha_a \geq \beta_a$ and that $\beta_a \geq 0 \land \gamma_a \geq 0$. The requirement on the relation between value of travel time ($\alpha_a$) and early schedule delay ($\beta_a$) is consistent with the plausible assumption that, when a flight is set to arrive early, the airline prefers landing over extending the trip by making a detour; the other requirement only states that values of schedule delay are not negative.

![Figure 2.1: Competitive no-toll equilibrium.](image)

With the untolled equilibrium characterized, we analyze the regulator’s problem of maximizing social welfare through a per-flight toll. First, consider the case of a time-invariant toll. As that toll does not vary over time, the airlines treat it as a constant operating cost and, for a given number of flights, it does not alter the scheduling decisions: the toll can only affect the number of flights. The regulator’s optimization problem follows:

$$\max_{SW} \int_0^{sf} (A - Bx)dx - \int_{ts}^{t_e} (K \cdot s \cdot C_p(t))dt - \int_{ts}^{t_e} (K \cdot c + K \cdot C_a(t))dt \quad (2.11)$$

---

16In fact, Lijesen (2006) finds evidence that $\gamma_p > \beta_p$, something that is usually found for road users.
where the first term is gross benefits for $sf$ travelers, the second is total passenger generalized costs (at $t$, a flow of $K$ flights will serve $s$ passengers each), and the third term is total airline costs that include constant and generalized costs (fares and tolls cancel out).

Rewriting,

$$SW = \int_0^{s^f} (A - Bx)dx - K \cdot \int_{t_s}^{t_e} (s \cdot C_p(t) + C_a(t))dt - K \cdot c \int_{t_s}^{t_e} dt$$

(2.12)

where the second equality uses that the duration of the peak is $f/K$, and that, in equilibrium, $s \cdot C_p(t) + C_a(t)$ is constant (condition in Eq. (2.8)).

Let $\tilde{\tau}$ be the time-invariant toll. Comparing the first-order conditions for welfare maximization and the airline zero-profit condition, we then obtain:

$$\frac{\partial SW}{\partial f} - \pi^e = s(A - B \cdot sf) - 2 \cdot \frac{\delta \cdot f}{K} - c - \left[ s(A - B \cdot sf) - c - \frac{\delta \cdot f}{K} - \tilde{\tau} \right]$$

(2.13)

As a consequence, the welfare maximizing time-invariant toll per flight is:

$$\tilde{\tau} = \frac{\delta \cdot f}{K}$$

(2.14)

This toll matches the flat toll for the road bottleneck (Arnott et al., 1993), because without altering the flights’ schedule, average (per flight) generalized costs are $\overline{\delta} \cdot f/K$, and the marginal social generalized cost is therefore $2 \cdot \overline{\delta} \cdot f/K$, which is fully consistent with the road case. As a consequence, it is straightforward that the second-best flat toll is the difference between the two. The flat-toll in Eq. (2.14) is equal to the marginal delay cost that a flight imposes on all airlines’ flights (including their passengers). This time-invariant toll induces an aggregate number of flights $f'$, which is second-best optimal, given that queuing is not eliminated. The fares will keep the dynamic structure that they have in the no-toll equilibrium (see Appendix 2.B for details), but the beginning and the end of the peak ($t_s$ and $t_e$) will be different, as the total number of flights is lower and the peak period shorter.

As queuing delay is a pure loss in this model, welfare can be improved further. The reason is that, any number of flights in an equilibrium with queues can be served in the same time interval, without queuing while incurring the same schedule delay costs. This requires an arrival rate equal to capacity throughout the peak, which cannot be achieved spontaneously in equilibrium as the flights closer to $t^*$ would incur a lower generalized cost. The first-best charge is the time-variant toll $\tau(t)$ that decentralizes this queue-free configuration. The toll is equal to the value of queuing delay per flight in the no-toll equilibrium, $\alpha \cdot T(t)$ in Figure 2.1, as it makes the sum of generalized cost and toll constant over time (the dynamic equilibrium condition) only when there is no queue and solely schedule delay costs are experienced. As travel delays are eliminated, the toll reduces the aggregate generalized cost by 50%, so that the marginal cost becomes equal to the generalized price. That is, the total cost becomes $\delta \cdot f^2/2K$, so the marginal cost
Chapter 2 Airport pricing in a dynamic bottleneck model of congestion

is \( \bar{\delta} \cdot f/K \), which is equal to the price faced by the first and last flight, and therewith by all flights. Denoting \( f^* \) as the optimal aggregate number of flights, the optimal toll is,

\[
\tau(t) = \frac{\bar{\delta} \cdot f^*}{K} - \begin{cases} 
\bar{\beta} \cdot (t^* - t) & \text{if } t < t^* \\
\gamma \cdot (t - t^*) & \text{if } t \geq t^*
\end{cases}
\]  

(2.15)

We call this toll structure the dynamic atomistic toll, in contrast to the second-best flat toll of this problem, in Eq. (2.14). With \( \tau(t) \), the flight that arrives at \( t^* \) does not experience delays and pays a toll equal to the marginal social cost. The first and last flight face a schedule delay equal to the marginal social generalized cost and therefore do not pay any toll. In addition, the part of the toll that reflects passenger valuation of delays is transferred to them through the fare to maintain passenger equilibrium. The fare will thus show a stronger time variation than in the no-toll equilibrium.

This section has extended the arguably most used model of dynamic congestion in the transport pricing literature to a case with a vertical structure, where passengers have a demand for trips, offered by atomistic, perfectly competitive carriers that make the scheduling decisions. In the rest of the chapter, we relax this assumption and allow for different degrees of market power through the analysis of various market structures.

2.2.2 The monopoly case

Here, we consider a market with a single airline facing a linear inverse demand as in Eq. (2.3). The monopoly carrier has as decision variables the number of flights \( F \) and how to schedule them, i.e. the time of departure of each flight. Let \( t_s \) be the time when the carrier schedules its first flight and \( t_e \) the time when the last flight is scheduled, then the airline’s profit is:

\[
\pi = \int_{t_s}^{t_e} K \cdot s \cdot \rho(t) - K \cdot c - K \cdot C_a(t) \, dt = K \int_{t_s}^{t_e} s[A - B \cdot sF] - sC_p(t) - c - C_a(t) \, dt
\]

\[
= s \cdot F \cdot [A - B \cdot sF] - F \cdot c - K \int_{t_s}^{t_e} s \cdot C_p(t) + C_a(t) \, dt
\]

(2.16)

The second equality uses Eq. (2.4), and the third equality, that the peak lasts \( F/K \).

We have shown that the last term on the right hand side of Eq. (2.16) reflects the road case with composite values of time \( \bar{\alpha}, \bar{\beta}, \) and \( \bar{\gamma} \). Since the airline faces no competition, the flights will be scheduled to minimize delays. The airline realizes that by choosing a departure rate equal to the runway capacity, it will achieve the minimum possible generalized cost, only facing schedule delay costs and no travel delay cost through queuing.

This allows us to write the generalized costs per flight as schedule delay costs that diminish linearly from \( \bar{\delta} \cdot F/K \) at \( t_s \) to zero at \( t^* \) and then grow to \( \bar{\delta} \cdot F/K \) at \( t_e \). Taking this into account, and considering a per-flight time-invariant toll \( \hat{\tau} \) (that is seen as parametric

\[17\] We are abstracting from potential entry in this setting, but we address this question in Section 2.3.
by the airline), the profit in Eq. (2.16) can be expressed as:

$$\pi = s \cdot F \cdot [A - B \cdot sF] - F \cdot c - \frac{\delta \cdot F^2}{2K} - F \cdot \hat{\tau}$$  \hspace{1cm} (2.17)$$

The airline first-order condition for profit maximization is,

$$\frac{\partial \pi}{\partial F} = s[A - B \cdot sF] - B \cdot s^2 F - c - \frac{\delta \cdot F}{K} - \hat{\tau} = 0$$  \hspace{1cm} (2.18)$$

which means that the (constant) full price paid by passengers is:

$$p = \rho(t) + C_p(t) = A - B \cdot sF = \frac{1}{s} \left( c + \frac{\delta \cdot F}{K} \right) + B \cdot sF + \frac{\hat{\tau}}{s}$$  \hspace{1cm} (2.19)$$

implying that the fare, that maintains equilibrium, is:

$$\rho(t) = \frac{1}{s} \left( c + \frac{\delta \cdot F}{K} \right) + B \cdot sF + \frac{\hat{\tau}}{s} - C_p(t)$$  \hspace{1cm} (2.20)$$

In contrast to the competitive case, this condition shows that the monopoly carrier charges to the passengers a markup of $B \cdot sF$. This is simply the number of passengers times the own-demand price sensitivity, the traditional market power effect first described, in the aviation context, by Pels and Verhoef (2004). Note that the fare is time-dependent, as the passengers’ generalized cost ($C_p(t)$) is not constant in this setting. Figure 2.2 depicts the time-invariant-toll equilibrium for a monopoly. There is no queue, and the first and last flight (at $t_s$ and $t_e$, respectively) experience a generalized cost of $\delta \cdot F/K$. The fulfillment of the first-order condition for profit maximization is represented in the vertical axis, where $s[A - B \cdot sF] = c + \delta \cdot F/K + B \cdot sF + \hat{\tau}$. The time-variant per-flight fare, $s \cdot \rho(t)$ in Eq. (2.20), is also depicted in Figure 2.2. The slopes of the passengers’ generalized cost ($C_p(t)$) and the airline’s delay cost ($C_a(t)$) are the same as in the optimum of the competitive case, and therefore the slope of the per-flight fare is also the same.

Now, the regulator’s maximization problem is:

$$SW = \int_{0}^{sF} (A - Bx)dx - K \cdot \int_{t_s}^{t_e} (s \cdot C_p(t) + C_a(t))dt - F \cdot c$$  \hspace{1cm} (2.21)$$

but, in contrast with the competitive case, the airline is scheduling the flights in such a way that there is no queue. Hence, the second term on the right-hand side of Eq. (2.21) is the same as derived in Eq. (2.17), shaping social welfare in the following way:

$$SW = \int_{0}^{sF} (A - Bx)dx - \frac{\delta \cdot F^2}{2K} - F \cdot c$$  \hspace{1cm} (2.22)$$

This is gross benefits of the $sF$ passengers minus total social costs; when there are no queuing delays, total generalized costs are $\delta \cdot F^2/2K$. Taking the derivative with respect
to $F$, we get the first-best condition:

$$s[A - B \cdot sF] = c + \delta \cdot \frac{F}{K} \tag{2.23}$$

At the optimum, full price equals marginal social cost, which is the sum of the marginal operating cost plus the marginal total generalized cost (including airlines and passengers through $\delta$). Comparing the monopolist’s first-order condition in Eq. (2.18) and the first-order condition for welfare maximization in Eq. (2.23), it is straightforward that the first-best toll is:

$$\hat{\tau} = -B \cdot s^2 F \tag{2.24}$$

The regulator corrects the market power exertion by subsidizing the airline, and does not have to give an incentive to the monopolist to internalize congestion. This subsidy ($-B \cdot sF$ per passenger) induces the optimal number of passengers, and is analogous to the one obtained in the static model (Pels and Verhoef, 2004). A monopoly airline internalizes all the congestion costs by scheduling the flights efficiently: there is no queuing and therefore there is no need for congestion pricing. In Figure 2.2, when the optimal subsidy is applied, the term $B \cdot s^2 F + \hat{\tau}$ disappears and the first-best condition in Eq. (2.23) is satisfied. Moreover, the per-flight fare ($s$ multiplied by the per-passenger fare) at the first and last flight is simply the airline’s cost per flight, as Figure 2.2 shows.

Despite the fact that in the monopoly case only a subsidy that decreases price is needed, it is important to emphasize that the dynamic atomistic toll in Eq. (2.15) could also be charged to the monopoly airline without altering social welfare, but transferring part of the

\[\text{We look at the full price of a flight (the full price of a trip $A - B \cdot sF$ times the number of passengers in a flight $s$) and the marginal social cost of a flight, but there is no loss of generality. The condition also implies that the full price of a trip equals the marginal social cost of a seat.}\]
monopoly carrier profits to the regulator. This is the result of the congestion technology: the monopoly airline cannot do better than setting the arrival rate equal to capacity in time windows where it has arrivals, so as to incur only schedule delay costs, regardless of the dynamic toll schedule it faces. To see this, it is enough to add the time-variant toll to the monopoly profit in Eq. (2.16):

$$\pi = s \cdot F[A - B \cdot sF] - F \cdot c - K \int_{t_a}^{t_e} s \cdot C_p(t) + C_a(t) + \tau(t) \, dt$$

(2.25)

By charging the dynamic atomistic toll in Eq. (2.14) to the monopoly, it is straightforward that \( s \cdot C_p(t) + C_a(t) \) in Eq. (2.25) cancels out with the time-variant part of the toll, and the airline will set full price of a flight equal to \( c \) plus the constant part of the toll per flight (\( \delta \cdot F^* / K \)) and the market power mark-up. By including the subsidy, the outcome will be the first-best.

Simultaneous competition between a small number of airlines in a Cournot fashion is, probably, the most studied market structure in the airport pricing literature, although mainly in the context of static congestion models. After having discussed monopoly and before moving on to the leader-fringe setting, it could have been the right place to discuss that setting; e.g., a duopoly. However, within the framework of the deterministic dynamic bottleneck model described above, with symmetric airlines, there is—in general—no arrival pattern equilibrium in pure strategies. For that reason the analysis is performed in a separate chapter. Chapter 3 provides a detailed analysis of the existence and uniqueness of equilibria in a setting of simultaneous scheduling by symmetric firms and it also extends the analysis to heterogeneous preferences.

2.3 A Stackelberg leader with a competitive fringe

We now turn to the case of competition between a Stackelberg leader and a follower that behaves competitively. This market structure was shown to be empirically relevant by Daniel (1995), and was studied further by Daniel and Harback (2008). They show that most of the U.S. airports have queuing patterns that are consistent with a stochastic bottleneck model, and exhibit evidence that the Stackelberg-fringe market structure is the one that fits best the observed queuing patterns. The purpose of this section is to assess the degree of internalization of congestion by the leader, and to derive the first-best tolls. To the best of our knowledge, there are two papers that study this type of set-up from a theoretical point of view. Brueckner and Van Dender (2008)—with a static congestion model—show that the internalization of self-imposed congestion by a Stackelberg leader facing a competitive follower can approach the atomistic levels, depending on the assumed substitution pattern, and that the first-best congestion toll can also approach the atomistic toll. On the other hand, Daniel (2009), with a dynamic bottleneck model of congestion, argues the need for atomistic tolls for both the leader and the competitive fringe with an analytical model that includes only the airlines, and therefore omits the vertical structure and the passengers’ role in the analysis.

The competitive follower can be interpreted as a group of competitive airlines, as in Section 2.2.1 with a free-entry condition. These airlines do not need to be small in general,
but only to have a small share of flights at the airport under consideration, where a single airline acts as a leader. Following the aviation literature, we use the term “fringe” for this group of airlines that behaves competitively, regardless of the temporal location of its flights. We assume that both the leader and the fringe treat the tolls that the regulator sets as parametric, and that when the Stackelberg leader makes its decisions, it is aware of the toll that the regulator applies to the fringe.\footnote{Brueckner and Verhoef (2010) point out that assuming that agents are large enough to exert market power and to recognize the impact of their decisions on overall congestion, but that they do not take into account the impact of their actions on the tolls, is a strong assumption. We maintain this assumption to focus on the first-order effects and comparison with earlier literature, but discuss how the solution proposed by Brueckner and Verhoef (2010) applies to our case in Section 2.3.3.}

### 2.3.1 Untolled equilibrium

To study the airlines’ interactions and assess the internalization of congestion, we first look at the no-toll equilibrium, following the framework proposed in Section 2.2. We extend the demand model to account for various substitution patterns between the leader and the fringe, by using the representative consumer model proposed by Dixit (1979). Demands are assumed to arise from the following strictly concave quadratic utility function:

\[
U(q_l, q_f) = A \cdot (q_l + q_f) - (B \cdot q_l^2 + 2 \cdot E \cdot q_l \cdot q_f + B \cdot q_f^2)/2,
\]

where \(A\), \(B\), and \(E\) are positive, and \(q_l\) and \(q_f\) are the number of passengers of the leader and the fringe respectively. This implicitly assumes that fringe carriers are perceived as perfect substitutes, and gives rise to the following inverse demand structure:

\[
D_i(q_i, q_j) = A - B \cdot q_i - E \cdot q_j \quad i \in \{l, f\} \land j \neq i
\]

(2.26)

where \(A\) represents the maximum reservation price, \(B\) is the own-demand sensitivity parameter, and \(E\) is the cross-demand sensitivity parameter. We assume \(B \geq E \geq 0\) in general, and usually \(B > E > 0\) so that outputs are imperfect substitutes. Perfect substitutability is a special case of our specification (\(E = B\)), while \(E = 0\) has airlines serving independent markets. This specification allows us to account for horizontal product differentiation that may come from particular aspects that may differ across carriers and make passengers perceive airlines as imperfect substitutes (e.g., food and language).\footnote{A model of (vertical) product differentiation where firms also choose quality would be more general, but it would divert attention from the implications of dynamic congestion on internalization and pricing.}

The passengers’ equilibrium condition, that stipulates that the marginal willingness to pay has to be equal to the per-trip generalized cost, implies the following fare:

\[
\rho_i(t) = A - B \cdot q_i - E \cdot q_j - C_p(t) \quad i \in \{l, f\} \land j \neq i
\]

(2.27)

This again implies that all carriers, in general, charge a fare that depends on the time of departure, as \(C_p(t)\) does.

In this game, each fringe carrier has the departure time of its flight as a decision variable, and the aggregate volume of the fringe is determined by the zero-profit condition. As in Section 2.2.1, in equilibrium, the generalized cost per flight \((s \cdot C_p(t) + C_a(t))\) must be constant in a period where the fringe operates (see Eq. (2.8)), otherwise a carrier will
2.3 A Stackelberg leader with a competitive fringe

have an incentive to reschedule its flight. Moreover, in absence of the leader, this can only be possible by queuing in the center of the peak, i.e. around the desired time of arrival $t^\star$, because it is where schedule delays are lower. In order to balance schedule delay costs and queuing delay costs, the queue must build up until $t^\star$ and, only then, start to dissipate until it disappears completely. As we describe in Section 2.2.1, and derive in Appendix 2.B, there is a unique aggregate pattern of departures that makes the generalized cost per flight constant over time, that will be the equilibrium pattern during the time-window where the fringe operates.

On the other hand, the Stackelberg leader has as decision variables the number of flights and the departure time of each of its flights. Because it anticipates the behavior of the fringe, the leader’s timing best response can be reduced to scheduling flights joining the queue of the fringe operators, and/or to schedule flights outside this congested period—in the peak shoulders—with a departure rate equal to capacity and bearing only schedule delay costs.$^{21}$ This is because the fringe carriers have the same (composite) values of time as the leader, and therefore the leader cannot benefit from taking over the center by causing higher queuing delays in this period. As a consequence, its best scheduling strategy for its flights in the center is to join the fringe’s queue without exceeding the duration being $(f + l_c)/K$, and the first and last flight having the same generalized cost.$^{22}$

The fringe zero-profit equilibrium condition is then given by:

$$s \left[ A - B \cdot s f - E \cdot s(l_c + l_s) \right] - c - \frac{\bar{\delta} \cdot (f + l_c)}{K} = 0 \tag{2.28}$$

This condition defines $f$ as a function of $l_c$ and $l_s$, and, therefore, it defines the fringe’s response to a change in the number of flights set by the leader in both the center and the peak. The fringe’s number of flights depends not only on the number of flights set by the leader in the peak center ($l_c$), but also on those in the shoulder ($l_s$), unless $E = 0$. This is

$^{21}$The leader can set the departure rate equal to the capacity of the bottleneck in the peak shoulders and achieve the minimum time costs, because it does not face competition or potential entry in the peak shoulders.

$^{22}$Denote $t_{c1}$ the beginning of the center and $t_{c2}$ the end. The conditions that determine the generalized cost per flight are: $\bar{\beta} \cdot (t^\star - t_{c1}) = \bar{\tau} \cdot (t_{c2} - t^\star) \land t_{c2} - t_{c1} = (f + l_c)/K$. Solving for $t_{c1}$ and $t_{c2}$, the costs at the borders will be $((\bar{\beta} \cdot \bar{\tau})/\bar{\beta} \cdot (\bar{\beta} + \bar{\tau})) \cdot (f + l_c)/K$, and denoting $\bar{\delta} = (\bar{\beta} \cdot \bar{\tau})/\bar{\beta} \cdot (\bar{\beta} + \bar{\tau})$ we get the result above. This also implies that a fraction $\bar{\tau}/\bar{\beta} \cdot (\bar{\beta} + \bar{\tau})$ of the flights will arrive early (between $t_{c1}$ and $t^\star$) and a fraction $\bar{\beta}/\bar{\beta} \cdot (\bar{\beta} + \bar{\tau})$ of the flights will arrive late (between $t^\star$ and $t_{c2}$). The aggregate departure rate is obtained by equalizing the time-derivative of $s \cdot C_p(t) + C_a(t)$ to zero.
an important point to stress, because it allows us to identify the condition that makes the
fringe care only about what happens in the center. The latter is the assumption made by
Daniel (2009). The full independence case of our model ($E = 0$) is thus the case where
results may be comparable with Daniel’s (2009) findings.

Straightforward calculations (see Appendix 2.C for all derivations) yield the following
conditions:

$$-1 \leq \frac{\partial f}{\partial l_c} < \frac{\partial f}{\partial l_s} \leq 0$$

which imply that the leader anticipates that any reduction in quantities (through a re-
duction in frequency either in the center or in the shoulders) will be met by an increase
in the fringe’s number of passengers, or, equivalently, new entry, until the fringe profit is
again zero.

There are two effects driving the fringe’s response. First, as airlines are perceived as
(imperfect) substitutes, any reduction in output by the leader will induce a shift in the
inverse demand of the fringe, that will induce an increase in the fringe’s output (this can
be seen in the first term of Eq. (2.28)). Second, the airlines are imposing congestion
on each other, and the leader predicts that any frequency reduction is partially offset by
an increase of the number of flights set by the follower in response to reduced queuing
(the third term in Eq. (2.28)). The substitutability effect is the same for changes in the
number of flights in the center and in the shoulders, but the congestion effect happens only
in the center, where there is congestion interaction. This is the reason why the fringe’s
response is stronger for changes in the center than in the shoulders ($\frac{\partial f}{\partial l_c} < \frac{\partial f}{\partial l_s}$).

When products are perfect substitutes ($E = B$), $\frac{\partial f}{\partial l_c} = -1$, which means that any
change in the leader’s number of flights in the center is fully offset by an opposite change
of equal magnitude by the fringe. This is because the zero-profit condition in Eq. (2.28),
determines a unique value for the aggregate number of flights when airlines are perfect
substitutes. In the other extreme, when airlines serve independent markets ($E = 0$), the
substitution effect disappears and only the congestion effect survives. This implies that
$\frac{\partial f}{\partial l_s} = 0 \land -1 < \frac{\partial f}{\partial l_c} < 0$, because in the shoulders there is no congestion. The
general case of imperfect substitutability ($B > E > 0$) is, naturally, in between the two
cases above, satisfying Eq. (2.29) with strict inequalities.

With the response of the fringe defined, we can look at the first-order conditions for the
Stackelberg leader and derive the equilibrium. In this untolled equilibrium, the leader’s
profit can be separated into two terms, the profit from the operations in the peak center
and the profit from the peak shoulders. In the center, because of the fringe’s presence, the
generalized cost per flight must be constant and equal to $\delta \cdot \frac{(f + l_c)}{K}$. In the shoulders,
the leader’s timing best response is to set the arrival rate equal to the bottleneck’s capacity and
experience only schedule delay costs. Since the duration of the entire peak has to be total
number of flights over capacity, $(f + l_c + l_s)/K$, the schedule delay cost of the first and last
flight is $\delta \cdot \frac{(f + l_c + l_s)}{K}$. The reason is the same as for the peak center; equalizing schedule
delay costs at the borders and knowing the duration, provide the necessary conditions to
determine costs at the borders (see footnote 22). Finally, with linear schedule delay costs,
the average generalized cost per flight in the peak shoulders will be the average between
the schedule delay cost at the exterior border of the shoulder (of the first and last flight)
and the schedule delay cost at the interior border of the shoulder (at the beginning and
2.3 A Stackelberg leader with a competitive fringe

end of the center): $\delta \cdot (f + l_c + l_s)/K + \delta \cdot (f + l_c)/K]/2 = \delta \cdot (f + l_c)/K + \delta \cdot l_s/(2K)$. This shapes the profit in the following way:

$$
\Pi = l_c \cdot \left( s[A - B \cdot s(l_c + l_s) - E \cdot sf] - c - \frac{\delta \cdot (f + l_c)}{K} \right) + l_s \cdot \left( s[A - B \cdot s(l_c + l_s) - E \cdot sf] - c - \frac{\delta \cdot (f + l_c)}{K} - \frac{\delta \cdot l_s}{2K} \right) \quad (2.30)
$$

The leader’s profit is a function only of $l_s$ and $l_c$, because we are already taking into account the equilibrium strategy in departure times. Any amount of flights the leader sets in the shoulders, $l_s$, will be scheduled at a departure rate equal to capacity, and any number of flights in the center $l_c$, will be scheduled such that generalized costs are constant (taking into account that the fringe also schedules flights in the center). The first-order conditions, in Appendix 2.C, show that the leader exerts market power through a markup that is less than the traditional monopoly markup, because of the fringe’s offsetting behavior. We also find that the fraction of flights that the leader sets in the shoulders is:

$$
\frac{l_s}{l_s + l_c} = 1 - \frac{Es^2 + \bar{\delta}/K}{Bs^2 + \bar{\delta}/K} < 1 \quad (2.31)
$$

This shows that in the untolled equilibrium the leader always schedules flights in the peak center, a key result to understand the scheduling behavior of the leader. Recall that in our model there are two sources of inefficiency: the number of flights and the timing of flights. As explained above, the flights in the shoulders are scheduled without queuing and, in this sense, efficiently; on the other hand, the flights in the center are scheduled inefficiently as they share the queuing pattern with the competitive fringe. Therefore, the fraction $l_s/(l_s + l_c)$ is a direct measure of the leader’s degree of efficiency in timing. As this ratio is always below 1, the leader never behaves fully efficiently in terms of timing.

When demand is imperfectly elastic and airlines are imperfect substitutes (i.e. $0 < E < B$), the leader schedules flights in both the peak center as well as in the peak shoulders ($l_s/(l_s + l_c) > 0$), behaving partially inefficient in terms of scheduling. This is also the case when the outputs are independent ($E = 0$). In the case of perfect substitution ($E = B$) and when demand is perfectly elastic ($B = E = 0$), the leader sets all of its flights in the peak center (so the peak center occupies the full peak), queuing along with the fringe, and being fully inefficient. The reason is that the leader knows that the fringe reacts to increases in $l_c$ by offsetting them, so that the fringe will make room for the leader’s flights; and conversely, if the leader decreases the number of flights in the peak center by shifting to the shoulders, the fringe will increase the number of flights raising the generalized costs. When the fringe fully offsets the changes in the leader’s number of flights, the leader is better off setting all the flights in the peak center along with the fringe. When this effect is partial, the leader is better off setting part of the flights in the center.

The implications for congestion internalization are now straightforward to identify. The leader fails to fully internalize self-imposed congestion because the offsetting behavior of the fringe (shown in Eq. (2.29)) reduces its incentives to decrease output. When demands are perfectly elastic or when demand is imperfectly elastic and products are perfect substi-
tutes, the leader does not internalize any congestion and behaves atomistically (consistent with its own demand becoming, in practice, perfectly elastic). In the case of full independence and imperfect substitutability, the leader internalizes only a fraction of the self-imposed congestion, because the offsetting behavior of the fringe is partial. These results reproduce previous findings, regarding internalization of self-imposed congestion in a Stackelberg-fringe competition, by Brueckner and Van Dender (2008), but now in a dynamic congestion model. Our result for full independence is also similar to the result by Daniel (2009), who finds that the leader sets a fraction of the flights in the peak center that ranges from 0 to 1 in the untolled equilibrium (Daniel’s proposition 1). Daniel argues that the leader sets all of the flights in the peak shoulders when the number of flights by the fringe is fixed. This is also true in our model, and is obtained when market are independents and only the fringe faces a perfectly inelastic demand.

2.3.2 First-best tolls

Tolls are required to correct the inefficiency present in the two margins of choice of the airlines: number of flights and trip timing. The number of flights is not optimal as a result of market power exertion by the leader and the lack of full internalization of congestion by all airlines. The inefficiency in trip timing is due to the fact that queuing is a pure loss in this model: any amount of flights queuing in a certain period of time can be rescheduled to arrive during the same interval in such a way that there is no queuing while schedule delays do not increase, therefore reducing social costs.

As a result of this, the optimal timing decision by carriers must satisfy an aggregate departure rate equal to capacity, so that there are no queuing delays and no spare capacity. This allows us to drop the differentiation between the leader’s flights in the center and shoulders, because there is no center with queuing. Let \( l \) be the number of flights of the leader and \( f \) the fringe’s number of flights. Because, in the first-best optimum, the first and last flight must experience the same cost (only schedule delay cost) and the duration of the peak is \((l + f)/K\), delay costs will equal \( \delta \cdot (l + f)/K \) in the borders, and they will decrease linearly to zero at \( t^* \). This yields a total social delay cost of \( (l + f)\cdot \delta \cdot (l + f)/2K \), and the first-best conditions, equating marginal social cost to full price for both the leader and the fringe, are given by:

\[
s[A - B \cdot sl - E \cdot sf] = s[A - B \cdot sf - E \cdot sl] = c + \frac{\delta \cdot (l + f)}{K} \tag{2.32}
\]

Denote \( f^* \) and \( l^* \) the first-best number of flights that solve Eq. (2.32). As the fringe does not exert market power, it is inefficient only in the timing decisions (excessive queuing). This implies that the congestion toll that has to be charged to the fringe is the dynamic atomistic toll described in Section 2.2.1:

\[
\tau(t) = \frac{\delta \cdot (f^* + l^*)}{K} - \begin{cases} 
\beta \cdot (t^* - t) & \text{if } t < t^* \\
\gamma \cdot (t - t^*) & \text{if } t \geq t^*
\end{cases} \tag{2.33}
\]

This toll is the marginal social cost of the first-best equilibrium (first term on the RHS of Eq. (2.33)) minus the schedule delay cost at time \( t \). It gives the incentive to each
2.3 A Stackelberg leader with a competitive fringe

fringe carrier to schedule its single flight such that the aggregate departure rate equals capacity, because it is the only timing equilibrium that yields a constant generalized price over time (the experienced schedule delay cancels out with the time-variant part of the toll). A higher departure rate would generate queuing delays and, therefore, a higher and unbalanced generalized price over time. A lower aggregate rate will generate room for new entry of fringe carriers, that would occur until there is no spare capacity.

To derive the optimal toll that the leader has to pay, we need to derive its best response in timing and number of flights, when the fringe faces the dynamic atomistic toll in Eq. (2.33). With static model of congestion, this is straightforward, as the only decision variable is the number of flights. In the present setting, the leader also chooses the departure time of each of its flights. The main result of our analysis is that the first-best congestion toll for the leader is not unique. There is a time-invariant toll that can be charged to the leader in order to achieve the first-best outcome, a time-variant toll that also yields the efficient outcome, but there are also various other pricing schemes.

- **Time-invariant toll**

The Stackelberg leader has the potential to schedule its flights without incurring queuing delays, as we discussed in Section 2.2.2 for a monopoly, but it has reduced incentives to do so in the no-toll equilibrium of this game because of the fringe’s presence. However, when the regulator imposes the dynamic atomistic toll in Eq. (2.33) to the fringe, the leader realizes that it can schedule flights efficiently (without queuing and operating at capacity), and knows that this keeps the fringe completely out of its own period of operation. This is because the fringe, when facing the atomistic toll in Eq. (2.33), experience a constant generalized price per flight equal to the marginal social cost ($\delta \cdot (f^* + l^*)/K$), regardless of the time of operation, as long as the aggregate departure rate is, at most, equal to capacity. In any other case, the generalized price will be higher. Thus, the leader, by setting its departure rate equal to capacity, effectively prevents entry from the fringe to its period, because the fringe is always better off operating at times when departures do not exceed the capacity.

The leader realizes that it is better off operating in the peak center, around $t^*$, where the schedule delays are lower. For any amount of flights $l$, the profit maximizing timing strategy is to set the departure rate equal to capacity (to have only schedule delay costs and prevent the fringe from entering), from $t_1$ to $t_2$ such that the cost of the first flight and the last flight is the same ($\bar{\delta} \cdot (t^* - t_1) = \bar{\gamma} \cdot (t_2 - t^*)$). As $t_2 - t_1 = l/K$, the delay cost at the borders equals $\bar{\delta} \cdot l/K$, and as schedule delay costs are linear and zero at $t^*$, the average delay cost per flight will be $\bar{\delta} \cdot l/2K$. Then, the leader’s profit and first-order condition, when facing a time invariant toll $\hat{\tau}$, is:

$$\Pi = l \cdot \left( s[A - B \cdot sl - E \cdot sf] - c - \bar{\delta} \cdot l/2K - \hat{\tau} \right)$$

$$\frac{\partial \Pi}{\partial l} = 0 \Rightarrow s[A - B \cdot sl - E \cdot sf] = c + \frac{\bar{\delta} \cdot l}{K} + (B + E \cdot \frac{\partial f}{\partial l}) \cdot s^2 l + \hat{\tau} \quad (2.34)$$

$^{23}$The derivation of the profit function is analogous to the monopoly case (see Eqs. (2.16) and (2.17)).
Chapter 2 Airport pricing in a dynamic bottleneck model of congestion

The leader fails to take into account the delays imposed on the fringe ($\delta \cdot f / K$ is not in the full price), and exerts market power (third term on the RHS of Eq. (2.34)), which in this case is reduced compared to the monopolistic case, because of the (partial) offsetting behavior of the fringe. The fringe’s full price depends on $l$ only through the demand side (the substitutability effect) because there is no queuing interaction. Hence, any reduction of frequency by the leader will result in a lower full price for the fringe, that—because of the free-entry (zero-profit condition)—translates into an output expansion by the fringe. The first-best flat-toll for the leader is simply the toll that corrects market power and congestion effects, and makes the full price set by the leader (RHS of Eq. (2.34)) equal to the marginal social cost (RHS of Eq. (2.32)):

$$\hat{\tau} = \frac{\delta \cdot f^*}{K} - (B + E \cdot \frac{\partial f}{\partial l}) \cdot s^2 l^*$$

(2.35)

This first-best toll consists of a market power subsidy (second term on the RHS), that depends on the substitution pattern, and a congestion charge that is independent of the amount of internalization of the untolled equilibrium. Our result shows that the congestion side of the toll appears to be different from what has been found in the literature before. Brueckner and Van Dender (2008) find that the leader should pay a congestion toll that lies in between the congestion imposed on the fringe and total marginal congestion costs (depending on the substitution pattern). We find that, when the regulator charges the dynamic atomistic toll to the fringe, because of the sequential nature of the game and the congestion technology, the leader does not fail to internalize self-imposed congestion. As a consequence, the regulator can induce the first-best outcome by charging the delays imposed by the leader on the fringe (analogous to the so-called “Cournot” toll).

When airlines are perfect substitutes ($E = B$), there is no need for market power subsidy as the fringe fully offset any reduction of flights ($\partial f / \partial l = -1$). As both types of agent behave atomistically, only the aggregate number of flights is defined and, therefore, the toll $\delta \cdot f^*/K$ will define the proportion of flights set by the leader and the fringe. In particular, the regulator can set the leader’s congestion toll to zero, meaning that the leader will supply the optimal output making the optimal timing decisions. In the case of full independence, the fringe does not exhibit the offsetting behavior as there is no substitutability effect nor congestion effect ($\partial f / \partial l = 0$), and the congestion part of the toll is uniquely determined because $f^*$ is unique (see Eq. (2.32)). In the general case of imperfect substitution ($0 < E < B$), the market power subsidy is lower because of the partial offsetting behavior of the fringe, and approaches zero as airlines become closer substitutes, while the congestion part of the toll remains uniquely defined by Eq. (2.32).

Figure 2.3 shows the (first-best) equilibrium that results from charging $\hat{\tau}$ in Eq. (2.35) to the leader, and the dynamic atomistic toll in Eq. (2.33) to the fringe. The leader schedules its flights to arrive in the center, between $[t_1, t_2]$, the fringe operates outside, between $[t_s, t_1]$ and $[t_1, t_e]$, the first-best conditions in Eq. (2.32) are satisfied, and there are no queuing delays. The leader charges a fare that depends on the time of departure ($\rho(t)$), its profit is equal to the saved queuing costs $\delta l^*^2 / 2K$ and the revenues from the market power effect ($l^* \cdot (B + E \cdot \partial f / \partial l) \cdot s^2 l^*$, not shown graphically). The congestion toll revenues (before subtracting the subsidy) per unit of capacity are equal to the shaded
2.3 A Stackelberg leader with a competitive fringe

area in Figure 2.3: the sum of the revenues from the leader (the rectangle in the center) and from the fringe (the two triangles at the shoulders).

\[ s[A - B \cdot sl^* - E \cdot sf^*] = s[A - B \cdot sf^* - E \cdot sl^*] \]

Figure 2.3: First-best equilibrium with the time-invariant toll charged to the leader.

- Time-variant toll

In this section we show that the first-best can also be attained by charging the dynamic atomistic toll in Eq. (2.33) to both the leader and the fringe, if, in addition, the market power subsidy is given to the leader. To see this, consider a leader’s flight that is scheduled to arrive at \( t \) to the destination. The profit that the leader gets from that flight is given by:

\[
\pi(t) = s \cdot [A - B \cdot sl - E \cdot sf - C_p(t)] - [C_a(t) + c] - [\tau(t)]
\]

(2.36)

where the first term in brackets is the fare that the leader charges for that flight (marginal willingness to pay minus passengers’ generalized cost), the second bracketed term is the airline’s cost from operating that flight, and the last term is the dynamic atomistic toll in Eq. (2.33). The negative component of the profit that depends on the time of arrival, that will determine the best response in timing, is the generalized cost per flight minus the time dependent part of the toll:

\[
[s \cdot C_p(t) + C_a(t) - \beta \cdot (t^* - t) \text{ if } t \leq t^*] - [\gamma \cdot (t - t^*) \text{ if } t \geq t^*]
\]

(2.37)

For any number of flights by the leader, the timing decisions must minimize the sum of the costs in Eq. (2.37) for all flights, and the unique way to do so is to schedule them such that the departure rate does not exceed the capacity of the bottleneck. This is because it minimizes the generalized cost per flight (first term in Eq. (2.37)), that will consist of only schedule delay costs and, as a consequence, the time varying component of the profit.
Chapter 2 Airport pricing in a dynamic bottleneck model of congestion

will be zero (the schedule delay costs cancel out with the time-varying part of the toll). This is also compatible with the competitive fringe’s reaction, as fringe carriers facing the dynamic atomistic toll never schedule flights to exceed capacity. As the best response in timing makes the profit per flight constant \( s \cdot C_p(t) + C_a(t) + \tau(t) = \bar{\delta} \cdot (f^* + l^*) / K \), the leader’s total profit can be written as the number of flights \( l \) times the profit per flight:

\[
\Pi = l \cdot \left( s[A - B \cdot sl - E \cdot sf] - c - \frac{\bar{\delta} \cdot (f^* + l^*)}{K} \right)
\]  

(2.38)

The first-order condition gives the full price set by the leader:

\[
\frac{\partial \Pi}{\partial l} = 0 \Rightarrow s[A - B \cdot sl - E \cdot sf] = c + \frac{\bar{\delta} \cdot (f^* + l^*)}{K} + (B + E \cdot \frac{\partial f}{\partial l}) \cdot s^2 l
\]  

(2.39)

This shows that the dynamic atomistic toll charged to the leader solves the externality inefficiency and only the market power exertion needs to be corrected with the same subsidy as in Eq. (2.35):

\[(B + E \cdot \frac{\partial f}{\partial l}) \cdot s^2 l^* \cdot \]

This result shows again that the congestion part of the first-best toll is not related with the degree of internalization in the untolled equilibrium. In this case, the dynamic atomistic toll for all carriers (leader and fringe) solves the externality inefficiency. This result also has an important implication for the financial situation of the airport. As Arnott et al. (1993) demonstrate, the self-financing results of Mohring and Harwitz (1962) for capacity investments hold for the bottleneck model with elastic demand. As we have shown, the results of our analysis parallel results for the road case regarding the toll; therefore, the self-financing result also holds when the toll in Eq. (2.33) is charged to both groups of airlines. If there are constant returns to scale in capacity provision, the revenues from the first-best toll then exactly cover the costs of providing the optimal capacity.24 This differs from earlier results because first-best tolls are now not discounted by the fraction of congestion that is internalized by carriers; therefore, the self-financing result is not overturned by the internalization of congestion. However, the market-power subsidy does upset exact self-financing under neutral scale economies. This is because under marginal cost pricing and constant returns to scale, the surplus will be zero. When part of the revenues is used to subsidize the firm with market power, there will be insufficient revenue to cover capacity costs. The shortfall in self-financing equals the aggregate airlines’ profit under constant returns to scale. With this time-varying pricing scheme, the leader makes a lower profit compared to the time-invariant case (this time only from the ability to exert market power), and revenues are equally higher. In fact, this is the pricing scheme that yields the highest revenue for the regulator.

- Alternative schemes

In addition to the two different ways to deal with the congestion inefficiency of the leader described above, namely a flat toll equal to the delays imposed on all the fringe’s flights or the dynamic atomistic toll, there are other pricing schemes that can induce the

\[24\text{In general, “the ratio of the revenue collected from the optimal toll to the costs of constructing optimal capacity equals the elasticity of construction cost with respect to capacity” (Arnott et al., 1993).}\]
2.3 A Stackelberg leader with a competitive fringe

first-best outcome. Although the market power distortion has to be corrected in any case with the subsidy in Eq. (2.35), \((B + E \cdot \delta f / \partial l) \cdot s^2 l^*\), the regulator can induce the leader to set the full price of its flights equal to the marginal social cost with regimes that correct the congestion effect differently.

First, note that if the leader is forced to give up a time-window \([t_1, t_2]\) around \(t^*\), of duration \(f^*/K\), and with \(t_1\) and \(t_2\) satisfying the condition of equal schedule delay costs \((\bar{\beta} \cdot (t^* - t_1) = \bar{\gamma} \cdot (t_2 - t^*))\), the first-best can again be attained (given that the market power distortion is being corrected). In this case the fringe will operate around \(t^*\), between \([t_1, t_2]\), paying the dynamic atomistic toll, ensuring that full price equals marginal social cost. The generalized cost per flight at \(t_1\) and \(t_2\) equals \(\bar{\delta} \cdot f^*/K\). The leader’s best response is to schedule its flights with a departure rate equal to capacity outside the “forbidden period” \([t_1, t_2]\), and such that the first and last flight experience the same generalized cost. As a result, it schedules \(l^*\) flights from \(t_s\) to \(t_1\) and from \(t_2\) to \(t_e\), without queuing, earning the saved queuing costs as profit. The leader’s first and last flight incur a generalized cost (per flight) of \(\bar{\delta} \cdot (f^* + l^*)/K\), hence satisfying the first-best condition in Eq. (2.32). The leader has no incentives to schedule more (nor less) flights, because the marginal revenue of the first and last flight is exactly equal to the marginal cost \((s \cdot \rho(t_e) - C_a(t_e), c = s \cdot \rho(t_s) - C_a(t_s)) = 0\) in Figure 2.3).

The toll regime that induces this outcome is the dynamic atomistic toll in Eq. (2.33) for the fringe, and for the leader the per-flight market power subsidy to correct dead-weight losses, and a toll arbitrarily higher than \(\bar{\delta} \cdot (f^* + l^*)/K\) only during the period \([t_1, t_2]\). The latter works as a barrier for the leader to operate in the peak center, as it makes him better off operating outside it, not paying the toll. In fact, this is equivalent to restrict the interval of time where the leader can operate.

This configuration is similar to the previous time-invariant toll setting in the sense that the full price does not change, because the gain in costs by the fringe (resulting from operating closer to \(t^*\)) is offset by higher tolls, and the cost increase of the leader is offset by the absence of congestion tolls. This makes this setting identical to the time-invariant case analyzed above in social welfare, consumer surplus, profit per firm (hence total profit) and total revenue. The difference, besides the times of operation for each firm, is that the tax revenues are not the same for each individual firm, but total tax revenues remain unchanged. In fact, there is a continuum of configurations, where the leader faces a time restriction (or barrier-toll) and a flat toll, that follows these properties. These configurations are defined by more elaborate patterns of temporal separation of leader and fringe operations, and the congestion tolls become a more complicated matter.

Furthermore, there is also a continuum of alternative schemes that deal with the leader’s congestion inefficiency with different time-varying tolls. This is due to the fact that the leader, knowing that the fringe faces the dynamic atomistic toll, will never schedule its flights in a way that causes the aggregate departure rate to exceed capacity. When the atomistic toll is charged to the leader, the schedule delay cost cancels out with the time-varying part of the toll at any time (see Eq. (2.37)). Thus, the marginal cost (to the firm) of a flight is only the fixed part of the toll, that is set equal to the marginal social cost by the regulator. If the regulator charges to the leader a toll with a time-varying component

\[\tau(t) = \bar{\delta} \cdot (f^* + l^*)/K\]

This equilibrium is not shown graphically, but it is enough to see Figure 2.3 and change \(l^*\) for \(f^*\) (and vice versa). The duration of the center is \(t_2 - t_1 = f^*/K\).
that is lower than the experienced schedule delay cost (i.e. less steep than the atomistic toll), they will no longer cancel out; as a result, the marginal cost will be the fixed part of the toll plus the time-varying cost. As the latter is now above zero, the former has to be reduced (with respect to the atomistic toll) in order to make the sum equal to marginal social cost. These toll schedules would produce toll revenues that are in between those from the flat toll, and those from the time varying atomistic toll.

2.3.3 Manipulable tolls

We assume in the analysis above that an airline, that is large enough to exert market power and to recognize the impact of its decisions on overall congestion (and followers), does not take into account the impact of its actions on the tolls. We are aware that this is a strong assumption, but it is common to most previous works. Brueckner and Verhoef (2010) propose a manipulable toll rule, designed to induce the social optimum when carriers predict the impact of their decisions on tolls, that can also be applied to our problem. They propose an adjustment such that the carriers’ profit plus the (manipulable) toll liability varies perfectly in parallel with social surplus. In our problem, the welfare maximizing tolls can be straightforwardly adjusted with their methodology. For example, consider the toll regime where the leader pays the time-invariant toll in Eq. (2.35); the adjusted first-best time-invariant toll rule, designed to be “manipulated” by the leader, is given by:

\[
T(l) = \frac{3 f^2}{2 \cdot K} \cdot \frac{1}{\partial f / \partial l} - (B + E \cdot \frac{\partial f}{\partial l}) \cdot \frac{s^2 \cdot l^2}{2} + T_c
\]

(2.40)

where \(T_c\) is a constant and \(\partial f / \partial l\) is independent of \(l\), and it is obtained in a way analogous to the one in Appendix 2.C for the untolled setting. By charging this toll rule, a leader that anticipates the effect of his decisions on a parametric toll (as in Eq. (2.35)) will have a pricing strategy that will lead to a social welfare maximizing number of flights and fares. This is because the marginal change in profit, from an increase in the number of flights, that is due to the toll rule \((\partial T(l) / \partial l)\) is exactly the parametric toll in Eq. (2.35) when evaluated at the social optimum.

The adjustment for the time-variant tolling regime follows the same logic: the market power subsidy has to be the same as above (second term on the right-hand side of Eq. (2.40)); the time-invariant component of the congestion toll rule has to be adjusted in a similar way as above (in this case including a term involving the own number of flights); and the time-variant component of the congestion toll rule needs to have the same slope as before, i.e. to vary over time perfectly in line with schedule delay costs.

2.4 Conclusions

This chapter studies airlines’ interactions and scheduling behavior, together with airport efficient pricing, using a deterministic bottleneck model of congestion. We confirm that an airline acting as a Stackelberg leader, facing a competitive group of fringe carriers, partially internalizes self-imposed congestion in the sense that, without facing tolls, it schedules fewer flights than perfectly competitive carriers would, achieving lower social congestion.
costs. Consistent with findings in the earlier literature using static models of congestion (e.g., Brueckner and Van Dender, 2008), the degree of internalization of self-imposed congestion depends critically on the assumed demand substitution pattern. Nevertheless, and what is new, our results suggests that social welfare maximizing congestion tolls do not depend crucially on the degree of internalization, and that the time-variant tolls derived for perfectly competitive carriers apply also to a monopoly airline and to a setting where a Stackelberg leader interacts with a group of competitive carriers as followers.

Our analysis suggests that optimal congestion pricing may have a more significant role than what has been suggested in the earlier literature based on static models. Moreover, the efficient fully time-variant congestion toll regime results in a revenue for the airport that restores the well known self-financing result for congested facilities Mohring and Harwitz (1962). Still, if the market power distortion is corrected with a subsidy drawn from the airport’s budget, the self-financing result is upset. Our results also suggest that the political feasibility of congestion pricing would be enhanced compared to earlier studies, as efficient congestion charges do not vary with market shares, and therefore are less likely to be perceived as inequitable.

We also find agreement with Daniel (1995, 2009) and Daniel and Harback (2008) in that dynamic atomistic tolls are efficient in markets well represented by an interaction between a leader and a competitive fringe as the follower, but we show that this is not the only efficient solution. The non-uniqueness of social welfare maximizing congestion tolls in this setting allows for other pricing schemes that also achieve the social optimum.

Incorporating heterogeneity and studying step-tolling are natural extensions of the present analysis, to complement Daniel’s (2009) work. Our model allows for the inclusion of heterogeneity in values of time and preferences for both airlines and passengers. Certainly, the equilibrium and optimal toll will depend on the type of heterogeneity considered. Step-tolling, a relevant alternative in practice, may bring important benefits compared with the social optimum; as the number of steps is increased, it approaches the dynamic atomistic congestion toll, and, consequently, also its efficiency and consumer surplus increases, approaching the optimal values (see van den Berg (2012)). Finally, the analysis of simultaneous competition between airlines with market power is also a natural extension of this analysis.
Appendix 2.A  Glossary of notation

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>Inverse demand intercept (reservation price)</td>
</tr>
<tr>
<td>$B$</td>
<td>Inverse demand own-quantity sensitivity parameter</td>
</tr>
<tr>
<td>$c$</td>
<td>Marginal cost per flight</td>
</tr>
<tr>
<td>$C_g(t)$</td>
<td>Generalized cost of arriving at $t$, for an agent type $g$</td>
</tr>
<tr>
<td>$d(\cdot)$</td>
<td>Inverse demand function</td>
</tr>
<tr>
<td>$E$</td>
<td>Inverse demand cross-quantity sensitivity parameter</td>
</tr>
<tr>
<td>$f$</td>
<td>Aggregate number of flights scheduled by perfectly competitive airlines</td>
</tr>
<tr>
<td>$F$</td>
<td>Aggregate number of flights scheduled by a monopoly airline</td>
</tr>
<tr>
<td>$g$</td>
<td>Agent type: $p$ for passengers and $a$ for airlines</td>
</tr>
<tr>
<td>$K$</td>
<td>Capacity of the bottleneck</td>
</tr>
<tr>
<td>$l$</td>
<td>Aggregate number of flights scheduled by the Stackelberg leader</td>
</tr>
<tr>
<td>$p_i$</td>
<td>Full price for a passenger traveling with airline $i$</td>
</tr>
<tr>
<td>$q_i$</td>
<td>Number of passengers traveling with airline $i$</td>
</tr>
<tr>
<td>$Q(t)$</td>
<td>Length of the queue at time $t$</td>
</tr>
<tr>
<td>$\dot{Q}$</td>
<td>Rate of change of the queue length</td>
</tr>
<tr>
<td>$r_d$</td>
<td>Aggregate departure rate</td>
</tr>
<tr>
<td>$s$</td>
<td>Number of passengers per aircraft</td>
</tr>
<tr>
<td>$t$</td>
<td>Time of arrival</td>
</tr>
<tr>
<td>$t_d$</td>
<td>Time of departure</td>
</tr>
<tr>
<td>$t_e$</td>
<td>Last moment of operation (moment of the last arrival)</td>
</tr>
<tr>
<td>$t_s$</td>
<td>First moment of operation (moment of the first arrival)</td>
</tr>
<tr>
<td>$t^*$</td>
<td>Desired arrival time</td>
</tr>
<tr>
<td>$T(t)$</td>
<td>Travel delay of a vehicle arriving at time $t$</td>
</tr>
<tr>
<td>$\alpha_g$</td>
<td>Value of travel time for agent type $g$</td>
</tr>
<tr>
<td>$\beta_g$</td>
<td>Value of early schedule delay for agent type $g$</td>
</tr>
<tr>
<td>$\gamma_g$</td>
<td>Value of late schedule delay for agent type $g$</td>
</tr>
<tr>
<td>$\rho_i(t)$</td>
<td>Fare charged by airline $i$ for a flight arriving at $t$</td>
</tr>
<tr>
<td>$\tau(t)$</td>
<td>Time-variant per-flight toll</td>
</tr>
<tr>
<td>$\hat{\tau}$</td>
<td>Time-invariant per-flight toll</td>
</tr>
</tbody>
</table>

Table 2.1: Glossary of notation.

Appendix 2.B  Derivation of the equilibrium in the perfect competitive case

2.B.1  Equilibrium in scheduling

This Section determines the unique equilibrium values for the beginning ($t_s$) and the end ($t_e$) of the peak period, as well as the departure rate function $r(t)$ that defines the
2.B Derivation of the equilibrium in the perfect competitive case

equilibrium (aggregate) queuing pattern. The travel delay, \( T(t) \), and queue length, \( Q(t) \), of a flight that departs at \( t \) are given by:

\[
T(t) = \frac{Q(t)}{K} \quad \text{and} \quad Q(t) = \int_{\hat{t}}^{t} (r(u) - K) du
\]

(2.41)

where \( \hat{t} \) is the most recent time at which there was no queue. Let \( \tilde{t} \) be the departure time for an on-time arrival \( (\tilde{t} + T(\tilde{t}) = t^*) \), and consider a flight that departs at \( t \) and arrives early \( (t < \tilde{t}) \). The generalized cost of that flight is:

\[
s \cdot C_p(t) + C_a(t) = \overline{\alpha} \cdot T(t) + \beta \cdot (t^* - t - T(t))
\]

(2.42)

The equilibrium condition states that the generalized cost per flight has to be constant over time. By equating the time-derivative of Eq. (2.42) to zero, we obtain the equilibrium departure rate for early arrivals:

\[
\frac{d[s \cdot C_p(t) + C_a(t)]}{dt} = \overline{\alpha} \cdot \left( \frac{r(t)}{K} - 1 \right) - \beta \cdot \left( 1 + \overline{\alpha} \cdot \left( \frac{r(t)}{K} - 1 \right) \right) = 0
\]

\[\Rightarrow r(t) = \frac{K \cdot \overline{\alpha}}{\overline{\alpha} - \beta} \quad \forall \ t \in [t_s, \tilde{t}) \]

(2.43)

Analogous calculations give the following equilibrium departure rate for late arrivals:

\[
r(t) = \frac{K \cdot \overline{\alpha}}{\overline{\alpha} + \gamma} \quad \forall \ t \in [\tilde{t}, t_e]
\]

(2.44)

Using that the first and last flight must experience the same generalized cost in equilibrium, and that the peak duration is \( f/K \), the start and end of the peak period can be derived, together with the equilibrium generalized cost per flight:

\[
\beta \cdot (t^* - t_s) = \gamma \cdot (t_e - t^*) \quad \land \quad t_s - t_e = f/K
\]

\[\Rightarrow t_s = t^* - \frac{\gamma}{\beta + \gamma} \cdot \frac{f}{K} \quad \land \quad t_e = t^* + \frac{\beta}{\beta + \gamma} \cdot \frac{f}{K} \quad \forall \ t \in [t_s, t_e]
\]

(2.45)

Finally, straightforward calculations yield the departure time for an on-time arrival:

\[
\hat{t} = t^* - \frac{\beta}{\overline{\alpha}} \cdot \frac{\gamma}{\beta + \gamma} \cdot \frac{f}{K}
\]

(2.46)

As it can be seen above, the conditions that we need to impose are \( \overline{\alpha} > \beta > 0 \) and \( \gamma > 0 \), so that the variables have the correct sign. As the empirical literature suggests (Morrison and Winston (1989); Lijesen (2006)), the values of time for passengers satisfy these conditions \( (\alpha_p > \beta_p > 0 \text{ and } \gamma_p > 0) \), and, as a consequence, the results hold also
Chapter 2 Airport pricing in a dynamic bottleneck model of congestion

when airlines’ do not incur schedule delay costs ($\beta_a = 0$ and $\gamma_a = 0$).

2.B.2 Equilibrium fare variation over time

With the equilibrium rates described, we can study how the fare $\rho(t)$ changes over time by using Eqs. (2.2), (2.10), and taking the derivative with respect to $t$:

$$\frac{\partial \rho(t)}{\partial t} = -\frac{\partial C_p(t)}{\partial t} = -\alpha_p \frac{\partial T(t)}{\partial t} - \begin{cases} \beta_p \frac{\partial (t^* - t)}{\partial t} & \text{if } t < t^* \\ \gamma_p \frac{\partial (t - t^*)}{\partial t} & \text{if } t \geq t^* \end{cases}$$

(2.49)

where we use that, in equilibrium, queuing delays, $T(t)$, have a slope of $\beta/\alpha$ for early arrivals and $-\gamma/\alpha$ for late arrivals. This reveals that only when the ratios $\alpha_p/\beta_p$ and $\alpha_p/\gamma_p$ equal the airlines’ willingness to accept delays in order to reduce travel times ($\alpha_a/\beta_a$ and $\alpha_a/\gamma_a$). On the other hand, when the passengers’ ratios $\alpha_p/\beta_p$ and $\alpha_p/\gamma_p$ are lower (higher) than the airlines’ ratios, the fare will be higher (lower) for passengers traveling closer to $t^*$.

Appendix 2.C Fringe’s response and leader’s first-order conditions

From Eq. (2.28), we can solve for $f$ and take the derivative with respect to $l_c$ and $l_s$ in order to obtain the fringe’s reaction to changes in number of flights by the leader. Solving for $f$, we obtain:

$$f = s \left[ A - E \cdot s(l_c + l_s) \right] - c - \frac{\delta \cdot l_c}{K}$$

(2.51)

Taking the derivative of Eq. (2.51) with respect to $l_c$, we find the response of the fringe to a change in the number of flights that the leader schedules in the peak center:

$$\frac{\partial f}{\partial l_c} = -\frac{E s^2}{B s^2 + \delta/K} - \frac{\delta/K}{B s^2 + \delta/K} = -\frac{E s^2 + \delta/K}{B s^2 + \delta/K} \equiv \phi$$

(2.52)

Since $E < B$, it follows that $-1 < \phi < 0$. That is, a frequency change by the leader in the peak center, yields an opposite change in number of flights by the fringe, but that is not equal in magnitude because of the assumed substitution pattern in demand. In the case where outputs are perfect substitutes ($E = B$), $\phi = -1$, which means that any frequency
2.C Fringe’s response and leader’s first-order conditions

reduction by the leader in the congested period is fully offset by an increase in number of flights by the competitive fringe. When outputs are independent $(E = 0)$, the response of the follower still partially offsets a leader’s frequency change.

Differentiating Eq. (2.51) with respect to $l_s$ gives the response of the fringe to a change in the number of flights scheduled in the peak shoulders:

$$\frac{\partial f}{\partial l_s} = - \frac{E s^2}{B s^2 + \delta/K} \equiv \lambda > \phi$$

(2.53)

When $0 \leq E < B$, the response of the fringe, to an increase of the leader number of flights scheduled in the peak shoulders, satisfies $-1 < \lambda \leq 0$. Note that $\lambda > \phi$ means that the response $\phi$ is stronger than the response $\lambda$, because both are negative.

With these expressions, we can derive the first-order conditions for profit maximization. Taking the derivatives of the profit in Eq. (2.30), we get the following:

$$\frac{\partial \Pi}{\partial l_c} = 0 = s[A - B \cdot s(l_c + l_s) - E \cdot sf] - c - \frac{\delta \cdot (f + l_c)}{K}$$

$$- [(B + \phi E) \cdot s^2(l_c + l_s)] - \left[\frac{\delta \cdot (l_c + l_s)}{K} \cdot (1 + \phi)\right]$$

(2.54)

$$\frac{\partial \Pi}{\partial l_s} = 0 = s[A - B \cdot s(l_c + l_s) - E \cdot sf] - c - \frac{\delta \cdot (f + l_c + l_s)}{K}$$

$$- [(B + \lambda E) \cdot s^2(l_c + l_s)] - \left[\frac{\delta \cdot (l_c + l_s)}{K} \cdot \lambda\right]$$

(2.55)

In both first-order conditions, the last two terms in square brackets on the right-hand side show the market power markup and the reduced incentives to internalize self-imposed congestion, respectively. By subtracting Eqs. (2.54) and (2.55), we can explicitly write the fraction of flights that the leader schedules in the shoulders:

$$\frac{\partial \Pi}{\partial l_c} - \frac{\partial \Pi}{\partial l_s} = \frac{\delta \cdot l_s}{K} - [\phi E \cdot s^2(l_c + l_s)] + [\lambda E \cdot s^2(l_c + l_s)] - \frac{\delta \cdot (l_c + l_s)}{K} \cdot (1 + \phi - \lambda) = 0$$

$$\Rightarrow \frac{l_s}{l_s + l_c} = \frac{E \cdot s^2(\phi - \lambda) + \delta/K(1 + (\phi - \lambda))}{\delta/K} = 1 - \frac{E s^2 + \delta/K}{B s^2 + \delta/K} \quad (2.56)$$
Chapter 3

On the existence and uniqueness of equilibrium in the bottleneck model with atomic users
3.1 Introduction

Individual users of roads and other transportation facilities are usually assumed to be small in the sense that they control a negligible fraction of total traffic. Yet large users are prevalent in many settings. Commercial airlines and rail companies often account for a sizable fraction of total traffic at airports and on rail networks stations. Postal services and major freight shippers operate large vehicle fleets that travel long distances each day. For example, FedEx handles about 150 daily flights out of Memphis International Airport and its air-cargo operations support tens of thousands of jobs. UPS operates on a similar scale out of Louisville, Kentucky (The Economist, 2013). Major employers such as government departments and large corporations can add substantially to traffic on certain roads at peak times. Large users such as these suffer from the congestion delays their own aircraft, trains, trucks, or other vehicles impose on each other. Thus, at airports, on rail networks, on congested roads, and on other transportation infrastructure networks, one would expect large users to internalize their self-imposed delays, and therefore to make different trip-related decisions than small users controlling the same aggregate traffic.

Following the terminology of game theory we will refer to small users as non-atomic, and large users that control a positive fraction of traffic as atomic. This terminology contrasts with the terminology used in the literature on airport congestion, beginning with Daniel (1995), in which users that control a negligible fraction of traffic and treat the congestion levels as parametric are called atomistic users. Somewhat confusingly, atomistic users are therefore non-atomic, and non-atomistic users are atomic.

There are several branches of literature on congestion with atomic users. In the aviation literature, Daniel (1995) was the first to recognize that airlines with market power and large shares of total traffic could internalize the delays their aircraft impose on each other. Brueckner (2002) showed that under Cournot competition airlines fully internalize self-imposed congestion. Further contributions in this line have been made by Pels and Verhoef (2004), Brueckner (2005), Zhang and Zhang (2006), Basso and Zhang (2007a), Brueckner and Van Dender (2008), and Silva and Verhoef (2013). In the context of road transportation, route-choice decisions by atomic users have been studied (e.g., Devarajan, 1981; Marcotte, 1987; Harker, 1988; Catoni and Pallottino, 1991; Miller et al., 1991; Cominetti et al., 2009; de Palma and Engelson, 2012). There is also a literature in operations research and computer science on atomic congestion games (e.g., Fotakis et al., 2008; Hoefer and Skopalik, 2009).

The above-mentioned studies have used static models of congestion except for Daniel (1995) who uses a stochastic queuing model empirically. Except for a few studies described below, traffic congestion with atomic users has not been studied with deterministic dynamic models. This is surprising because the timing of trips matters a great deal for both passenger and freight transportation, and congestion is largely a consequence of peak-period loads.

An earlier version of this chapter has been published in the Tinbergen Institute Discussion Paper series (Silva et al., 2014a). I am very grateful to Erik Verhoef for his constructive and incisive comments which have significantly improved the chapter. I would also like to thank participants at the Kuhmo Nectar Conference on Transportation Economics 2013 at Northwestern University. The main variables and parameters that are used in this chapter are summarized in Appendix 3.A.

---

26 An earlier version of this chapter has been published in the Tinbergen Institute Discussion Paper series (Silva et al., 2014a). I am very grateful to Erik Verhoef for his constructive and incisive comments which have significantly improved the chapter. I would also like to thank participants at the Kuhmo Nectar Conference on Transportation Economics 2013 at Northwestern University. The main variables and parameters that are used in this chapter are summarized in Appendix 3.A.
3.1 Introduction

The goal of this chapter is to investigate the fundamental questions of existence and uniqueness of equilibrium in trip-timing decisions with atomic users. To focus the analysis on fundamentals while minimizing mathematical complications we use Vickrey’s (1969) bottleneck model. The essence of the bottleneck model is that users trade off the costs of queuing delay at the bottleneck with the costs of schedule delay (i.e., arriving earlier or later than desired). The bottleneck model has been used to study many aspects of trip-timing decisions with congestion including: congestion pricing, route choice on simple road networks, mode choice, trip chaining, parking congestion, staggered work hours, and flextime. Existence and uniqueness of equilibrium in the bottleneck model have also been established under relatively general assumptions about trip-timing preferences and heterogeneity of non-atomic users (e.g., Newell, 1987; Lindsey, 2004). However, very little consideration has been given to atomic users in studies that use either the bottleneck model or other dynamic models.

A few studies have employed a variant of the bottleneck model in which time is discretized, and the number of users is finite, so that each user controls a positive measure of traffic (e.g., Levinson, 2005; Zou and Levinson, 2006; Otsubo and Rapoport, 2008; Werth et al., 2014). However, these studies assume that each user controls only one vehicle so that self-internalization of congestion does not come into play. To the best of our knowledge, only Daniel (2009) and Silva et al. (2014c) have explored the scheduling decisions of atomic agents in the standard, continuous-time bottleneck model. These studies consider, in the context of aviation, a sequential competition between a Stackelberg leader with market power and a group of perfectly competitive airlines (non-atomic users). Both studies show that, when users have homogeneous preferences, non-atomic users schedule all their flights during the peak period when passengers prefer to arrive. The Stackelberg leader schedules a fraction of its flights during the peak as well. Queuing time at the bottleneck evolves at the same rate as in the standard model of non-atomic players. The leader schedules its remaining flights earlier and later in the off-peak (and less popular) periods and limits its departure rate to bottleneck capacity so that no queue develops.

The existence of a unique equilibrium in Daniel (2009) and Silva et al. (2014c) hinges on the sequential nature of the game they consider, and the assumption that there is only one atomic agent. By contrast, we focus in this chapter on settings with two atomic users who make scheduling decisions simultaneously. The solution concept we employ is pure strategy Nash equilibrium (PSNE). We establish three major results. First, we show that a PSNE may not exist. We demonstrate this for an example featuring two identical atomic users who each control half of the total traffic. The trip-timing preferences of each vehicle in each fleet are described by parameters \(\{\alpha, \beta, \gamma, t^*\}\), where \(\alpha\) is the cost of travel time, \(\beta\) is the cost of schedule delay early, \(\gamma\) is the cost of schedule delay late, and \(t^*\) is the desired arrival time. We show that if \(\gamma > \alpha\), a PSNE does not exist.

Second, for the same example, we show that if \(\gamma \leq \alpha\) a multiplicity of PSNE exists in which no queuing occurs and the timing of departures is socially optimal. The PSNE differ according to the departure rates of individual users and the equilibrium costs they incur. Depending on parameter values, one of the two users can incur up to three quarters of total costs. The cases \(\gamma > \alpha\) and \(\gamma \leq \alpha\) are both of theoretical interest, and each may be relevant in particular settings. Most empirical studies of scheduling preferences for automobile drivers have obtained estimates that satisfy \(\gamma > \alpha\) (e.g., Small, 1982;
Chapter 3 Bottleneck model with atomic users

Wardman, 2001; Asensio and Matas, 2008). In addition, de Palma and Fontan (2001) list estimates from eleven studies and, including their own estimates, there are ten cases with \( \gamma > \alpha \), and two cases with \( \gamma < \alpha \). Daniel and Harback (2008) find that \( \gamma < \alpha \) holds for most airlines at major US airports.

Third, we consider a variant of the example in which the two users differ in their desired arrival times \( t^* \) and can have fleets of different size. We show that, independent of the relative size of \( \alpha \) and \( \gamma \), a multiplicity of PSNE can exist in which no queuing occurs but the timing of departures is not optimal. Depending on parameter values, the PSNE may begin earlier than, later than, or at the same time as the social optimum. The order in which users depart can be suboptimal as well. Nevertheless, by internalizing self-imposed congestion costs the two users realize much, and possibly all, of the potential cost savings from either centralized traffic control or time-varying congestion tolls.

These examples demonstrate that neither the existence of equilibrium nor the uniqueness of an equilibrium (if one exists) is guaranteed under conditions where a unique PSNE does exist if all the traffic were controlled by non-atomic users. Given the central role of equilibrium models in the analysis of transportation systems, these results are troubling and highlight the need for further research.

The chapter is organized as follows. Section 3.2 reviews the no-toll equilibrium and social optimum in the standard bottleneck model with non-atomic users. Section 3.3 demonstrates the possible non-existence of PSNE, and the non-uniqueness of individual departure rates and costs where a PSNE does exist. Section 3.4 demonstrates the possible nonuniqueness of PSNE in the timing of departures when no queuing occurs, and the degree of inefficiency relative to the social optimum. Section 3.5 concludes the chapter.

### 3.2 The bottleneck model with homogeneous non-atomic users

The bottleneck model was developed by Vickrey (1969) and extended by Arnott et al. (1990, 1993). It is reviewed in Arnott et al. (1998) and de Palma and Fosgerau (2011), and the summary here is brief. In the model, all users travel from a common origin to a common destination along a single link that has a bottleneck with fixed flow capacity, \( s \). Without loss of generality, travel times from the origin to the bottleneck and from the bottleneck to the destination are normalized to zero. If there is no queue upstream of the bottleneck, travel time through the bottleneck is also zero and departure time from the origin coincides with arrival time at the destination. If the departure rate exceeds \( s \), a queue develops. Let \( \hat{t} \) be the most recent time at which there was no queue, and \( r(t) \) the aggregate departure rate from the origin at time \( t \). The number of vehicles in the queue is then:

\[
Q(t) = \int_{\hat{t}}^{t} (r(u) - s)du.
\]

Travel time through the bottleneck is \( T(t) = Q(t)/s \), and a traveler who departs at time \( t \) arrives at time \( t_a = t + T(t) \).

Following Small (1982) and de Palma et al. (1983), users are assumed to have a desired arrival time \( t^* \). They incur a unit cost of \( \beta > 0 \) for arriving early, and a unit cost of \( \gamma > 0 \)
for arriving late. Travel time is valued at $\alpha$, with $\alpha > \beta$. The costs of schedule delay and travel time are additive so that the generalized cost of a trip, $c(t)$, is:

$$
c(t) = \alpha \cdot T(t) + \left\{ \begin{array}{ll}
\beta \cdot (t^* - t - T(t)), & t + T(t) \leq t^* \\
\gamma \cdot (t + T(t) - t^*), & t + T(t) \geq t^*
\end{array} \right.. \quad (3.1)
$$

The number (or measure) of users, $N$, is assumed to be exogenous (i.e., independent of trip cost). Each user decides when to depart from the origin by trading off schedule delay against travel delay. A pure strategy Nash equilibrium (PSNE) is a set of departure times such that no user can benefit (i.e., reduce trip cost) by unilaterally changing departure time while taking the departure times of all other users as given.

### 3.2.1 No-toll equilibrium

Let superscript $n$ denote the no-toll non-atomic PSNE, and $t^n_s$ and $t^n_e$ denote the start and end of the travel period. Let $\hat{t}$ be the departure time for which a user arrives on time (i.e., $\hat{t} + T(\hat{t}) = t^*$). In a PSNE with no toll, $c(t)$ must be constant during the travel period $[t^n_s, t^n_e]$. Users who arrive closer to $t^*$ must incur longer queuing delays in order to offset their lower schedule delay costs. The equilibrium aggregate departure rate is derived by differentiating Eq. (3.1) and setting the derivative to zero:

$$
r^n(t) = \left\{ \begin{array}{ll}
\frac{\alpha \cdot s}{\alpha - \beta}, & t \in (t^n_s, \hat{t}) \\
\frac{\alpha \cdot s}{\alpha + \gamma}, & t \in (\hat{t}, t^n_e)
\end{array} \right.. \quad (3.2)
$$

The assumption $\alpha > \beta$ assures that the departure rate for early arrivals is positive and finite. This condition is plausible since a user who is destined to arrive early is likely to prefer arriving early to prolonging the trip by making a detour. The condition is also supported by Small’s (1982) estimates for automobile commuting trips. For ease of reference we will sometimes call the departure rate for early arrivals the *early departure rate*, and the departure rate for late arrivals the *late departure rate*.

There are two further equilibrium conditions. One is that the first and last users to depart, who encounter no queue, must incur equal schedule delay costs:

$$
\beta \cdot (t^* - t^n_s) = \gamma \cdot (t^n_e - t^*). \quad (3.3)
$$

The other condition is that the travel period lasts for $N/s$:

$$
t^n_e - t^n_s = \frac{N}{s}. \quad (3.4)
$$

Together, equilibrium conditions (3.2), (3.3) and (3.4) yield:

$$
t^n_s = t^* - \frac{\gamma \cdot N}{\beta + \gamma \cdot s}, \quad (3.5)
$$
\[ t^n_e = t^* + \frac{\beta}{\beta + \gamma} \cdot \frac{N}{s}, \quad (3.6) \]
\[ \tilde{t} = t^* - \frac{\beta}{\alpha} \cdot \frac{\gamma}{\beta + \gamma} \cdot \frac{N}{s}, \]
\[ c^n(t) = \delta \cdot \frac{N}{s}, \quad t \in [t^n_s, t^n_e] \text{ with } \delta = \frac{\beta \cdot \gamma}{\beta + \gamma}. \]

Total costs are:
\[ TC^n = \delta \cdot \frac{N^2}{s}. \quad (3.7) \]

**3.2.2 Social optimum**

Queuing delay at the bottleneck is a deadweight loss. The social optimum therefore avoids queuing and minimizes total schedule delay costs. The departure rate is maintained at \( s \) over a continuous time interval chosen so that the first and last users incur the same schedule delay cost. The departure period is therefore the same as in the laissez-faire PSNE (cf. Eqs. (3.5) and (3.6)). Using superscript \( o \) (letter \( o \)) to denote the social optimum, these results are recorded for future reference as:
\[ r^o(t) = s, \quad t \in (t^o_s, t^o_e), \quad r^o(t) = 0 \text{ otherwise,} \quad (3.8) \]
\[ t^o_s = t^* - \frac{\gamma}{\beta + \gamma} \cdot \frac{N}{s}, \quad (3.9) \]
\[ t^o_e = t^* + \frac{\beta}{\beta + \gamma} \cdot \frac{N}{s}. \quad (3.10) \]

Total social costs are only half as large as in Eq. (3.7) for the no-toll equilibrium:
\[ TC^o = \frac{\delta}{2} \cdot \frac{N^2}{s} = \frac{1}{2} \cdot TC^n. \quad (3.11) \]

The difference between total costs in the no-toll equilibrium and social optimum, \( TC^n - TC^o \), serves as an upper bound on the benefits from self-internalization of congestion by atomic agents.

### 3.3 Existence and non-existence of equilibrium with homogeneous atomic users

In this section we study an example featuring two identical atomic users. We show that if \( \gamma > \alpha \), a PSNE in departure schedules does not exist. We then show that if \( \gamma \leq \alpha \), a PSNE does exist that entails no queuing and coincides with the social optimum. During early departures the two users can depart at somewhat different rates that add up to \( s \). If \( \gamma < \alpha \), their late departure rates can also differ. Moreover, with \( \gamma \leq \alpha \) the users can incur appreciably different total costs for their fleets. At the end of the section we briefly discuss how these results extend to more than two users.
Consider two atomic users, $A$ and $B$. Each user controls a fleet of $N/2$ vehicles. Trip-timing preferences for each vehicle are defined by the same $\{\alpha, \beta, \gamma, t^*\}$ parameter values. Thus, the total cost of a user’s fleet is simply the sum of the cost of each of its vehicles. Users $A$ and $B$ simultaneously choose departure schedules for their fleets. The schedule for user $i$ is a departure rate function, $r_i(t) \geq 0$. This function can be seen as a distribution function of the $N/2$ vehicles over some extended time interval such as a day. The function is not restricted to be continuous, and the possibility of mass departures will be considered. Each user recognizes that dispatching a vehicle at time $t$ may delay vehicles in its fleet that depart after $t$. A delay occurs if there is a queue at time $t$ that persists when the later vehicles depart. A delay also occurs if there is no queue prior to $t$, but the bottleneck is at capacity so that adding a vehicle to the departure schedule at $t$ creates a (small) queue.

As noted above, the existence of a PSNE in this example depends on whether $\gamma > \alpha$, or $\gamma \leq \alpha$. The two cases are considered in the following two subsections.

### 3.3.1 Non-existence of PSNE with $\gamma > \alpha$

When $\gamma > \alpha$, a PSNE does not exist. This result is formalized in the following proposition.

**Proposition 3.1.** If $\gamma > \alpha$, a PSNE in departure schedules does not exist.

We prove Proposition 3.1 in four steps. First, we prove that a PSNE without queuing does not exist (Lemma 3.1). Second, we prove that mass departures cannot arise in equilibrium (Lemma 3.2). Third, we show that there is a unique departure pattern with queuing such that a user cannot reduce its fleet costs by rescheduling a single vehicle (Lemma 3.3). Finally, we show that this departure pattern is not a PSNE because a user can reduce its fleet costs by rescheduling a positive measure of vehicles in the fleet (Lemma 3.4). This establishes that a PSNE with queuing does not exist either.

**Lemma 3.1.** When $\gamma > \alpha$, a PSNE without queuing does not exist.

**Proof:** Consider a pair of departure schedules, $\{r_A(\cdot), r_B(\cdot)\}$, such that no queuing occurs. Some vehicles must arrive late since otherwise a user could reduce its fleet costs by rescheduling some vehicles to just after $t^*$. Consider a period $(t_1, t_2)$ of late arrivals and assume that both users depart during this period (the case where only one user departs is considered later). The bottleneck must be used to capacity since otherwise a user could exploit the residual capacity by advancing departures for vehicles that are scheduled to depart later.

If user $i$ removes a vehicle from the departure schedule at time $t$ it saves a cost of

\[ C_i^-(t) = \gamma \cdot (t - t^*). \]  \hspace{1cm} (3.12)

Removing the vehicle saves the late-arrival cost incurred by the vehicle itself, but it has no effect on the rest of the fleet because there is no queue. If user $i$ instead adds a vehicle...

---

27This assumption seems realistic for delivery vans carrying merchandise or parcels to different customers. It may be inappropriate for vehicles in a military convoy or emergency vehicles traveling to an accident.
to the departure schedule, it increases its fleet costs by a marginal private cost (MPC) of

\[ C_i^+ (t) = \gamma \cdot (t - t^*) + \frac{\alpha + \gamma}{s} \cdot \int_t^\tau r_i (u) \, du, \]

(3.13)

where \( \tau \) is the time when the queue created by the additional vehicle disappears.

The first term on the right-hand side of Eq. (3.13) matches the right-hand side of Eq. (3.12). The second term is the delay cost imposed on user \( i \)'s other vehicles that depart from \( t \) to \( \tau \). Each of them suffers an increase in travel time of \( 1/s \) valued at \( \alpha \), and an increase in late arrival of \( 1/s \) valued at \( \gamma \).

The difference between the cost saved by removing a vehicle given by Eq. (3.12), and the cost of adding a vehicle to the same slot given by Eq. (3.13), arises when the bottleneck is at capacity but there is no queue. This asymmetry underlies the nonuniqueness of equilibrium considered later in the chapter.

Suppose user \( i \) advances the departure of a vehicle from \( t \) to \( t' \) where \( t_1 \leq t' < t \leq t_2 \). User \( i \)'s fleet costs change by:

\[
\Delta C_i = -C^-_i (t) + C^+_i (t') = -\gamma \cdot (t - t^*) + \gamma \cdot (t' - t^*) + \frac{\alpha + \gamma}{s} \int_t^{t'} r_i (u) \, du. \tag{3.14}
\]

The queue induced by adding the vehicle at \( t' \) vanishes at \( t \) because a departure time slot opened up at \( t \) when the vehicle was removed then. Let \( \lambda_{t,t'}^i = \int_t^{t'} r_i (u) \, du / (s \cdot (t - t')) \in [0, 1] \) be an auxiliary variable denoting the average fraction of capacity occupied by user \( i \) during the period \([t, t']\). The change in user \( i \)'s fleet costs can then be written as:

\[
\Delta C_i = \frac{\alpha + \gamma}{s} \cdot \lambda_{t,t'}^i \cdot s \cdot (t - t') - \gamma \cdot (t - t') = (\alpha + \gamma) \cdot \lambda_{t,t'}^i - \gamma \cdot (t - t').
\]

For a PSNE to exist, \( \Delta C_i \) must be nonnegative for both users which requires:

\[
\lambda_{t,t'}^i \geq \frac{\gamma}{\alpha + \gamma}, \quad t_1 \leq t' < t \leq t_2, \quad i = A, B. \tag{3.15}
\]

Since the bottleneck is fully utilized, \( \lambda_{t,t'}^A + \lambda_{t,t'}^B = 1 \). This condition is least restrictive if \( \lambda_{t,t'}^A = \lambda_{t,t'}^B = 1/2 \) in which case it reduces to \( \gamma \leq \alpha \) which is inconsistent with the assumption \( \gamma > \alpha \).

Now consider the possibility that only one user, say user \( A \), departs during \((t_1, t_2)\). This is not a PSNE if user \( B \) departs after \( t_2 \) since user \( B \) could reduce its costs by rescheduling some of its later vehicles into \((t_1, t_2)\). Doing so reduces their late-arrival costs, and the queue they create disappears during the time slots they vacated. Suppose user \( B \) does not depart after \( t_2 \). If \( t_1 = t^* \), user \( B \) does not depart late at all. But this cannot be a PSNE because user \( B \) could gain by rescheduling some of its early-arriving vehicles to just after \( t^* \), thereby reducing their schedule delay costs without imposing any delay on its other vehicles. If \( t_1 > t^* \), there must exist a late arrival period \((t_0, t_1)\) during which both users depart. But this case has already been considered, and shown to be inconsistent with a
PSNE when \( \gamma > \alpha \). \(QED\)

To this point we have assumed that users depart at a finite rate. In theory, a user could schedule a positive measure of vehicles to depart at a given moment. In practice, this might be achieved by assembling a convoy of vehicles on a link that has right-of-way over other links. Moreover, in the bottleneck model with non-atomic users and step tolls a PSNE may exist only if mass departures (of non-cooperating vehicles) are possible (Arnott et al., 1990; Lindsey et al., 2012). We now show that in the model with atomic users mass departures cannot occur in a PSNE, regardless of whether \( \gamma > \alpha \) or \( \gamma \leq \alpha \).

**Lemma 3.2.** A PSNE cannot exhibit mass departures.

**Proof:** See Appendix 3.B.

We now turn to the final possibility for a PSNE with \( \gamma > \alpha \) in which departure rates remain finite and queuing occurs.

**Lemma 3.3.** When \( \gamma > \alpha \), there is a unique departure pattern with queuing in which a user cannot reduce its fleet costs by rescheduling a single vehicle.

**Proof:** Consider a pair of departure schedules, \( \{r_A(\cdot), r_B(\cdot)\} \), and let \( \tilde{t} \) denote the departure time for which a user arrives on time (i.e. \( t = t^* - T(\tilde{t}) \)). Assume that a queue exists during a late-departure period \((t_l, t_{qe})\), where \( t_l \) is an arbitrary time that satisfies \( t_l > \tilde{t} \) and \( t_{qe} \) is the time when the queue disappears. Since users \( A \) and \( B \) are identical, it suffices to consider the best response of user \( A \) to \( r_B(\cdot) \). The MPC to user \( A \) of scheduling a vehicle at time \( t \in (t_l, t_{qe}) \) is

\[
C_A(t) = \alpha \cdot T(t) + \gamma \cdot (t + T(t) - t^*) + \frac{\alpha + \gamma}{s} \cdot \int_t^{t_{qe}} r_A(u) \, du. \tag{3.16}
\]

Eq. (3.16) has a similar interpretation to Eq. (3.13). The first two terms on the right-hand side comprise the cost incurred by the vehicle itself, and the third term is the delay cost imposed on user \( A \)'s other vehicles that depart from \( t \) to \( t_{qe} \).

User \( A \) could reschedule a vehicle from \( t \) to another time \( t' \in (t_l, t_{qe}) \). This would leave \( t_{qe} \) unchanged because the additional queuing time caused by inserting the vehicle at \( t' \) is offset by the reduction in queuing time due to removing the vehicle at \( t \). Eq. (3.16) therefore holds if \( t \) is replaced by any \( t' \in (t_l, t_{qe}) \). Hence, a necessary condition for \( r_A(\cdot) \) to be a best response to \( r_B(\cdot) \) is that \( C_A(t) \) in Eq. (3.16) is constant during the interval \((t_l, t_{qe})\). Differentiating Eq. (3.16) with respect to \( t \), and setting the derivative to zero, yields

\[
\frac{\partial C_A(t)}{\partial t} = \gamma + (\alpha + \gamma) \cdot \frac{\partial T(t)}{\partial t} - \frac{\alpha + \gamma}{s} \cdot r_A(t) = 0.
\]

Using the relationship \( \partial T(t)/\partial t = (r_A(t) + r_B(t) - s)/s \), this condition simplifies to

\[
\frac{\alpha \cdot s}{\alpha + \gamma} = r_B(t), \quad t \in (t_l, t_{qe}). \tag{3.17}
\]

According to Eq. (3.17), user \( A \) is willing to schedule a vehicle for late arrival when there is a queue if, and only if, user \( B \) is departing at exactly the rate \( \alpha \cdot s/(\alpha + \gamma) \).
This is none other than the equilibrium aggregate departure rate for the model with non-atomic users (cf. Eq. (3.2)). Eq. (3.17) also holds for user $B$ with $r_A(t)$ in place of $r_B(t)$. The aggregate departure rate during an interval of late arrivals must therefore be $r(t) = r_A(t) + r_B(t) = 2 \cdot \alpha \cdot s / (\alpha + \gamma)$. With $\gamma > \alpha$, $r(t) < s$ and the queue must be shrinking for late arrivals. Consequently, a queue must exist at time $\tilde{t}$ and it is possible to set $t_l = \tilde{t}$, where $\tilde{t} + T(\tilde{t}) = t^*$. This in turn implies

$$r_A(t) = r_B(t) = \frac{\alpha \cdot s}{\alpha + \gamma}, \quad t \in (\tilde{t}, t_{qe}).$$

Since a queue exists at time $\tilde{t}$, it must have built up during a period of early arrivals before $\tilde{t}$. Let $t_q$ be the time at which queuing begins. The MPC to user $A$ of scheduling a vehicle at any time $t \in (t_q, \tilde{t})$ is

$$C_A(t) = \alpha \cdot T(t) + \beta \cdot (t^* - t - T(t)) + \frac{\alpha - \beta}{s} \cdot \int_t^\tilde{t} r_A(u) \, du + \frac{\alpha + \gamma}{s} \cdot \int_\tilde{t}^{t_{qe}} r_A(u) \, du. \quad (3.18)$$

Again, the first two terms on the right-hand side of Eq. (3.18) comprise the cost borne by the vehicle itself. The third term is the cost imposed on user $A$’s other vehicles that depart after $t$ but still arrive early. Each of them suffers an increase in travel time of $1/s$ valued at $\alpha$, and benefits from a reduction in early arrival of $1/s$ valued at $\beta$. The last term in Eq. (3.18) is the cost imposed on user $A$’s other vehicles that arrive late.

A necessary condition for $r_A(\cdot)$ to be a best response to $r_B(\cdot)$ is for $C_A(t)$ to be constant during $(t_q, \tilde{t})$. Setting the derivative of $C_A(t)$ to zero, one obtains a counterpart to Eq. (3.17):

$$r_B(t) = \frac{\alpha \cdot s}{\alpha - \beta}, \quad t \in (t_q, \tilde{t}). \quad (3.19)$$

An analogous necessary condition applies for user $B$. Hence, the aggregate early departure rate must be $r(t) = r_A(t) + r_B(t) = 2 \cdot \alpha \cdot s / (\alpha - \beta)$.

In summary, the unique departure rate of the candidate PSNE during the full period of queuing is

$$r_A(t) = r_B(t) = \frac{r(t)}{2} = \begin{cases} \frac{\alpha \cdot s}{\alpha - \beta}, & t \in (t_q, \tilde{t}) \\ \frac{\alpha \cdot s}{\alpha + \gamma}, & t \in (\tilde{t}, t_{qe}) \end{cases}, \quad (3.20)$$

where

$$\tilde{t} + T(\tilde{t}) = t^*. \quad (3.21)$$

Because queuing begins at $t_q$, and ends at $t_{qe}$, cumulative departures during the period $(t_q, t_{qe})$ match cumulative arrivals:

$$\int_{t_q}^{t_{qe}} (r(u) - s) \, du = 0. \quad (3.22)$$

Eqs. (3.20), (3.21), and (3.22) define evolution of the queue for the candidate PSNE with queuing.
3.3 Existence and non-existence of equilibrium with homogeneous atomic users

By Lemma 3.1, no vehicles can depart without queuing after \( t_{qe} \), so departures end at time \( t_e = t_{qe} \). However, the cost of a vehicle trip at the beginning of the queuing period, \( c(t_q) \), is less than the cost at the end of the period, \( c(t_e) \), because the trip at \( t_q \) imposes a private delay cost on subsequent vehicles whereas the trip at \( t_e \) does not (if \( c(t_q) = c(t_e), C^+(t_q) > C^+(t_e) \) would hold and the user equilibrium condition would be violated). Departures must therefore occur during some time interval \([t_s, t_q)\) preceding \( t_q \). This interval is defined by two conditions. First, vehicle trip cost must be the same at \( t_s \) and \( t_e \) since otherwise a user could reschedule vehicles from the time with higher cost to the time with lower cost and reduce its overall fleet costs without causing any queuing. Second, the full departure period \([t_s, t_e]\) must be long enough for all \( N \) vehicles to pass the bottleneck. Eqs. (3.3) and (3.4) for the non-atomic PSNE therefore hold for the candidate PSNE:

\[
\beta \cdot (t^* - t_s) = \gamma \cdot (t_e - t^*),
\]

\[
t_e - t_s = \frac{N}{s}.
\]

The two users each control \( N/2 \) vehicles and schedule the same number of vehicles during the queuing period. Therefore, they must also schedule the same number during \([t_s, t_q)\):

\[
\int_{t_s}^{t_q} r_A(u) \, du = \int_{t_s}^{t_q} r_B(u) \, du,
\]

where

\[
r_A(u) + r_B(u) = s, \quad u \in (t_s, t_q).
\]

The early departure schedule defined by Eqs. (3.25) and (3.26) is consistent with a PSNE as far as trip timing by individual vehicles. The first vehicle scheduled at \( t_s \) creates the same MPC as all vehicles scheduled during \([t_q, t_e]\). Vehicles scheduled during \((t_s, t_q)\) create a lower MPC, so that rescheduling them to any time outside \((t_s, t_q)\) would increase total fleet costs. Rescheduling any vehicle into \((t_s, t_q)\) would also increase fleet costs because it would impose a queuing delay on all vehicles departing later until \( t_e \), and would therefore create a higher MPC than a vehicle departing at \( t_q \). QED

Together, Eqs. (3.20)–(3.26) define the candidate PSNE with queuing. Using superscript \( c \) to denote this candidate, the critical times are:

\[
t^c_s = t^* - \frac{\beta}{\beta + \gamma} \cdot \frac{N}{s},
\]

\[
t^c_e = t^* + \frac{\beta}{\beta + \gamma} \cdot \frac{N}{s},
\]

\[
\tilde{t}^c = t^* - \frac{2 \cdot \alpha \cdot (\gamma + \beta)}{\alpha + \beta} \cdot \frac{N}{s},
\]

\[
t^c_q = t^* - \frac{\alpha \cdot (\gamma - \alpha)}{(\alpha + \beta) \cdot (\beta + \gamma)} \cdot \frac{N}{s}.
\]

We now show that the candidate PSNE just derived is not a PSNE because either user can reduce its fleet costs by rescheduling a positive fraction of its vehicles. This result is
Chapter 3 Bottleneck model with atomic users

formalized as:

**Lemma 3.4.** When \( \gamma > \alpha \), the candidate departure pattern with queuing in which users cannot reduce costs by rescheduling a single vehicle is not a PSNE with respect to rescheduling a positive fraction of the fleet.

**Proof:** The candidate PSNE is depicted in Figure 3.1. Cumulative departures during the whole travel period are shown by the piecewise linear schedule \( OABC \). During the first interval \((t^c_s, t^c_q)\), the two users depart early at rates consistent with Eqs. (3.25). During the remaining interval \((t^c_q, t^c_e)\), user B contributes to cumulative departures the portion between schedule \( ADGE \) and the horizontal line \( AF \). User A contributes the equally big portion between schedule \( ABC \) and schedule \( ADGE \). During the interval \((t^c_q, \hat{t})\), each user departs early at rate \( \alpha \cdot s / (\alpha - \beta) \), and during \((\hat{t}, t^c_e)\) each user departs late at rate \( \alpha \cdot s / (\alpha + \gamma) \). At time \( \hat{t} \), each user’s cumulative departures over the interval \((t^c_q, \hat{t})\) match cumulative bottleneck throughput over the same interval.

![Figure 3.1: Candidate PSNE with \( \gamma > \alpha \) and queuing.](image)

Suppose user A reschedules all its vehicles from period \((t^c_q, \hat{t})\) to period \((\hat{t}, t^c_e)\), and maintains a departure rate of \( \gamma \cdot s / (\alpha + \gamma) \) during \((\hat{t}, t^c_e)\). As a consequence, the queue caused by B between \( t^c_q \) and \( \hat{t} \) fully dissipates at \( \hat{t} \). Since user B departs at rate \( \alpha \cdot s / (\alpha + \gamma) \) during \((\hat{t}, t^c_e)\), the total departure rate during this period equals \( s \) and the bottleneck is fully utilized without queuing. Moreover, since user B keeps the bottleneck fully utilized during \((\hat{t}, t^c_e)\), all vehicles in both fleets complete their trips by \( t^c_e \). Consequently, with the deviant schedule, all of user A’s vehicles can complete their trips within the same time period as in the candidate PSNE, but without queuing. It is straightforward to show (see
Appendix 3.C) that user A’s fleet incurs lower total costs with the deviant schedule than with the candidate PSNE schedule. Hence the candidate PSNE is not a PSNE. QED

The reader may wonder why the candidate PSNE is robust to deviations in which a user reschedules a single vehicle, but not robust to rescheduling part of the fleet. The reason is that the first type of deviation comprises a zero measure of traffic whereas the second type comprises a positive measure. Rescheduling a positive measure of traffic can increase or decrease the MPC imposed by other vehicles in the fleet, and a judicious, coordinated, rescheduling of vehicles can result in a net reduction in fleet costs.

Proposition 3.1 establishes that a PSNE does not exist when $\gamma > \alpha$. As noted in the introduction, most empirical studies of scheduling preferences for automobile drivers have obtained estimates that satisfy this inequality. However, there is little evidence either on the trip-timing preferences of users that may control large shares of road traffic (e.g., freight shippers), or on preferences for travel by other modes of transportation. Daniel and Harback (2008), without making explicit the role of the passengers’ valuation of time, estimate that $\gamma < \alpha$ holds for many US airlines. Thus, it is of interest to study the existence and nature of PSNE when $\gamma \leq \alpha$.

### 3.3.2 Existence and Nature of PSNE with $\gamma \leq \alpha$

In this section we establish two results for the case $\gamma \leq \alpha$. First, we show that there exists a unique PSNE in the aggregate departure schedule that coincides with the social optimum (Proposition 3.2). Second, we show that the two users' individual departure rates are not uniquely defined in the PSNE (Proposition 3.3), and the users can incur different fleet costs (Section 3.3.2.3).

#### 3.3.2.1 Existence of PSNE with $\gamma \leq \alpha$

Define $r_E \equiv \frac{2 \cdot \alpha \cdot (\beta + \gamma) + \gamma \cdot (\beta + 2 \cdot \gamma) - \sqrt{\beta^2 \cdot \gamma^2 + 4 \cdot \alpha \cdot (\beta + \gamma)^2 \cdot (\alpha + \gamma)}}{\gamma \cdot (\alpha + \gamma)} \cdot s, \ r_E \in (0, \frac{\gamma}{\alpha + \gamma} \cdot s)$.

**Proposition 3.2.** If $\gamma \leq \alpha$, there exists a PSNE in which the aggregate departure schedule coincides with the social optimum. During late arrivals, each user’s departure rate satisfies: $\frac{\gamma}{\alpha + \gamma} \cdot s \leq r_i(t) \leq \frac{\gamma}{\alpha + \gamma} \cdot s, \ t \in (t^*, t_0^i), \ i = A, B$. During early arrivals, a sufficient condition for a PSNE is that each user’s departure rate satisfies $r_i(t) \geq r_E, \ t \in (t_0^i, t^*)$, $i = A, B$.

**Proof:** The proof entails establishing four results: 1) A PSNE with queuing does not exist. 2) All PSNE without queuing must coincide with the socially optimal departure pattern given by Eqs. (3.8), (3.9), and (3.10). 3) Given the lower bound on individual departure rates for late arrivals, neither user can gain by rescheduling a single vehicle. 4) Given the lower bounds on individual departure rates for early and late arrivals, neither user can gain by rescheduling part of its fleet.

**Result 1.** A PSNE with queuing does not exist.

By Lemma 3.2, mass departures cannot be part of a PSNE. Any candidate PSNE with queuing must satisfy conditions (3.20), (3.21), and (3.22). With $\gamma \leq \alpha$, these conditions
cannot all be satisfied since the aggregate early departure rate exceeds capacity, and the aggregate late departure rate is no less than capacity. Hence any queue cannot dissipate while users are departing, which is inconsistent with a PSNE.

**Result 2.** The socially optimal departure pattern is the only possible PSNE in the aggregate departure schedule.

Departures must occur at rate $s$ over a connected time interval since otherwise either user could reduce its fleet costs by rescheduling vehicles into “gaps” in the departure schedule when bottleneck capacity is not fully utilized. The departure period, $[t_s, t_e]$, must be as given by Eqs. (3.9) and (3.10) since otherwise trip costs at $t_s$ and $t_e$ would differ, and at least one user could reduce its fleet costs by rescheduling vehicles from the higher-cost endpoint to the lower-cost endpoint.

**Result 3.** A user cannot gain by rescheduling a single vehicle if each user’s departure rate satisfies the conditions in Proposition 3.2.

In the candidate PSNE with a socially optimal aggregate departure pattern, there is no queuing but the bottleneck is used to capacity. As in the proof of Lemma 3.1, it is therefore necessary to distinguish between the cost saved by removing a vehicle from the departure schedule (which does not affect other vehicles’ costs) and the cost of adding a vehicle (which creates a queue unless the vehicle is added at $t^*_s$). The respective costs are:

$$C^-_i(t) = \begin{cases} 
\beta \cdot (t^* - t), & t \in [t^*_s, t^*_e] \\
\gamma \cdot (t - t^*), & t \in [t^*, t^*_e] 
\end{cases}$$

$$C^+_i(t) = \begin{cases} 
\beta \cdot (t^* - t) + \frac{\alpha - \beta}{s} \int_t^{t^*} r_i(u) \, du + \frac{\alpha + \gamma}{s} \int_{t^*_s}^{t^*} r_i(u) \, du, & t \in [t^*_s, t^*_e] \\
\gamma \cdot (t - t^*) + \frac{\alpha + \gamma}{s} \int_t^{t^*} r_i(u) \, du, & t \in [t^*, t^*_e] 
\end{cases}$$

A vehicle can be rescheduled in four ways: (i) late to late, (ii) late to early, (iii) early to late, and (iv) early to early. Consider each possibility in turn.

1. **Rescheduling late to late:** Rescheduling a late vehicle to a later time is never beneficial because the vehicle’s trip cost increases, and other vehicles do not gain. Suppose a vehicle is rescheduled earlier from $t$ to $t'$ where $t^* \leq t' < t$. The change in fleet costs is given by Eq. (3.14):

$$\Delta C_i = -C^-_i (t) + C^+_i (t') = -\gamma \cdot (t - t') + \frac{\alpha + \gamma}{s} \int_{t'}^t r_i(u) \, du$$

$$= \lambda_{t', t}^* \frac{\gamma}{\alpha + \gamma}$$

where $\frac{s}{\alpha + \gamma}$ means identical in sign. Given $r_i(t) \geq \gamma \cdot s / (\alpha + \gamma)$ for $t \in (t^*, t^*_e)$, $\lambda_{t', t}^* \geq \gamma / (\alpha + \gamma)$, $\Delta C_i \geq 0$, and the deviation is not beneficial.

2. **Rescheduling late to early:** Rescheduling a late vehicle to an early time is clearly inferior to rescheduling it to $t^*$ because the vehicle incurs an early-arrival cost and creates a queue for a longer period. But rescheduling it to $t^*$ is not beneficial as per case i.

3. **Rescheduling early to late:** The best option in this case is to reschedule a vehicle from $t^*_s$. However, the gain is the same as for rescheduling a vehicle from $t^*_e$, and this is
3.3 Existence and non-existence of equilibrium with homogeneous atomic users

not beneficial as per case i. Rescheduling early to late therefore cannot be beneficial.

iv. Rescheduling early to early: The best option in this case is to reschedule a vehicle from $t^e_a$ to $t^e$, but this, too, is not beneficial for the same reason as in case iii.

This establishes that the candidate PSNE in Proposition 3.2 is robust to deviations in which a single vehicle is rescheduled.

Result 4. A user cannot gain by rescheduling a positive measure of its fleet.

If user $A$ reschedules vehicles to depart outside $[t^e_a, t^e]$, its fleet costs necessarily increase since the rescheduled vehicles experience greater schedule delay costs without benefiting the rest of the fleet. If user $A$ instead reschedules vehicles to depart inside $[t^e_a, t^e]$, queuing occurs. The optimal departure rate in the presence of a queue was derived in Section 3.3.1 and, although with $\gamma \leq \alpha$ a PSNE cannot exhibit queuing, we need to check if a deviation from the candidate to a setting with queuing is cost reducing. For early arrivals, user $A$ is willing to depart at a positive and finite rate only if condition (3.19) is satisfied; i.e. $r_B(t) = \alpha \cdot s / (\alpha - \beta) > s$. Since user $B$ departs at a rate less than $s$ in the candidate PSNE, user $A$ is better off scheduling the vehicles later. For late arrivals, user $A$ is willing to depart at a positive and finite rate only if condition (3.17) is satisfied so that $r_B(t) = \alpha \cdot s / (\alpha + \gamma)$. Proposition 3.2 stipulates that $r_B(t) \leq \alpha \cdot s / (\alpha + \gamma)$. If $r_B(t) = \alpha \cdot s / (\alpha + \gamma)$, and therefore $r_A(t) = \gamma \cdot s / (\alpha + \gamma)$, user $A$ is indifferent between departing or not so that the deviation does not reduce its fleet costs. If $r_B(t) < \alpha \cdot s / (\alpha + \gamma)$, user $A$ is better off scheduling the vehicles later. The only remaining possibility for gainful deviation is one that involves mass departures. In Appendix 3.D we prove that a deviation with mass departures is not gainful if the lower bounds on early and late departure rates stated in Proposition 3.2 are both satisfied. QED

The intuition for Proposition 3.2 is as follows. The only way for user $A$ to gainfully deviate from the candidate PSNE by rescheduling a single vehicle is to advance its departure during the late arrival period. Doing so reduces the vehicle’s late-arrival cost, but imposes queuing delay on user $A$’s vehicles that depart later. The tradeoff is not worthwhile if enough vehicles in $A$’s fleet have yet to depart, and queuing is sufficiently costly relative to late arrival. The lower bound on the late departure rate stated in Proposition 3.2 assures this condition is met. If user $A$ deviates by rescheduling vehicles in mass departures and imposes queuing delays on its other vehicles, the same argument holds. The tradeoff is not worthwhile because the lower bound on the late departure rate ensures that there are enough vehicles in the fleet yet to depart that will be negatively affected.

To avoid delaying other vehicles in its fleet, user $A$ must advance departures for vehicles that participate in a mass. Since vehicles in the mass suffer queuing delay, user $A$ can benefit from such a deviation only if the vehicles’ schedule delay costs are reduced enough. This requires that the vehicles were departing late over a period longer than the time they take to pass the bottleneck when in the mass. This, in turn, is possible only if user $B$ occupies a large enough share of bottleneck capacity during the candidate PSNE. The lower bound on the late departure rate identified in Proposition 3.2 ensures that this condition is not met.

To see why a lower bound on the early departure rate is also required, suppose that user $A$ does not depart during some time interval $(t, t^e)$. User $A$ can then reschedule its vehicles departing in some interval $(t^e, t')$ by launching them in a mass at a time $t_m \in$
None of its fleet departing after \( t' \) will be delayed by the mass. If \( t_m \) is chosen to 
minimize the total schedule delay costs of vehicles in the mass, their total costs will fall. 
However, if user \( A \) is scheduling enough departures during \((t_m, t^*)\), the deviation will not 
be gainful since the mass departure either imposes queuing delays on its other vehicles, 
or it has to include vehicles that were arriving early, which will suffer higher early arrival 
costs and queuing costs. If the early departure rate is high enough, such a mass departure 
is unfavorable. Proposition 3.2 identifies a minimum early departure rate to guarantee 
this when the late departure rate is held fixed at its minimum value: \( \gamma \cdot s/(\alpha + \gamma) \).

### 3.3.2.2 Non-uniqueness of PNSE with \( \gamma \leq \alpha \)

Although the aggregate departure schedule in the PSNE described in Proposition 3.2 is 
unique when \( \gamma \leq \alpha \), many pairs of departure schedules \( \{r_A(\cdot), r_B(\cdot)\} \) are consistent with 
the aggregate pattern. Thus, the PSNE is not unique in terms of individual departure 
rates. This result is formalized in the following proposition:

**Proposition 3.3.** During the late departure period \( t \in (t^*, t^*_s) \), the two users depart at 
the same rate \( r_A(t) = r_B(t) = s/2 \) if \( \alpha = \gamma \), but they can depart at different rates if 
\( \gamma < \alpha \). During the early departure period \( t \in (t^*_s, t^*) \), a continuum of departure schedules 
\( \{r_A(\cdot), r_B(\cdot)\} \) are consistent with the PSNE.

**Proof:** By Proposition 3.2, for \( t \in (t^*, t^*_s) \), \( \gamma/(\alpha + \gamma) \leq r_A(t)/s \leq \alpha/(\alpha + \gamma) \) and 
\( r_B(t) = s - r_A(t) \). If \( \gamma = \alpha, \gamma/(\alpha + \gamma) = 1/2 \) and therefore \( r_A(t) = r_B(t) = s/2 \). If \( \gamma < \alpha \), 
there is a continuum of pairs of \( \{r_A(t), r_B(t)\} \) that satisfy the equilibrium condition while 
assuring that all vehicles depart during the interval \([t^*_s, t^*_c]\). For \( t \in (t^*_s, t^*) \), the constraint 
\( r_i(t) \geq r_E \) is less strict because \( r_E < s/2 \). Any pair of departure schedules \( \{r_A(\cdot), r_B(\cdot)\} \) 
that satisfies \( r_E \leq r_A(t) \leq s - r_E \) and \( r_B(t) = s - r_A(t) \) satisfies the aggregate PSNE 
condition. \( \text{QED} \)

The non-uniqueness of individual departure rates in a PSNE implies that, unlike in the 
model for identical non-atomic users, users can experience different fleet costs in a PSNE. 
This prospect is examined further in the next subsection.

### 3.3.2.3 Asymmetric equilibrium costs with \( \gamma \leq \alpha \)

Clearly, if users \( A \) and \( B \) depart at the same rate throughout the departure period they 
incur the same fleet costs. In addition, there are many asymmetric PSNE departure 
schedules that result in the same average schedule delay costs for the two users and hence 
the same fleet costs. However, there are also many asymmetric departure schedules that 
result in different fleet costs.

Figure 3.2 depicts an illustrative example of a PSNE in which user \( A \) incurs lower 
fleet costs than user \( B \). In the example, user \( A \)’s departures are concentrated near \( t^* \) 
so that user \( A \) has lower average schedule delay costs than \( B \). User \( A \)’s departure rate 
is shown by solid lines and user \( B \)’s by broken lines. During the interval \((t^*_s, t_{BA})\), user 
\( A \) departs at the minimum rate \( r_E \) defined in Proposition 3.2. During the next interval 
\((t_{BA}, t^*)\), user \( B \) departs at rate \( r_E \) and user \( A \) at the complementary rate \( s - r_E \). During 
the first part \((t^*, t_{AB})\) of the late-arrival period, user \( A \) departs at the maximum rate

58
consistent with a PSNE for user $B$, $\alpha \cdot s / (\alpha + \gamma)$. During the remaining part $(t_{AB}, t^e)$ of the departure period, user $A$ departs at the minimum rate consistent with a PSNE for itself, $\gamma \cdot s / (\alpha + \gamma)$. The transition times $t_{BA}$ and $t_{AB}$ are such that each user dispatches a total of $N/2$ vehicles. Expressed in terms of user $A$’s fleet, the requisite condition is

$$r_E \cdot (t_{BA} - t^e) + (s - r_E) \cdot (t^* - t_{BA}) + \frac{\alpha}{\alpha + \gamma} \cdot s \cdot (t_{AB} - t^*) + \frac{\gamma}{\alpha + \gamma} \cdot s \cdot (t^e - t_{AB}) = \frac{N}{2}. \quad (3.31)$$

Total costs in the PSNE, $TC^o$, are a given; i.e. independent of users $A$’s and $B$’s individual departure rates. The maximum difference between the users’ costs, $TC_B - TC_A$, can therefore be found by minimizing $TC_A$ with respect to $t_{BA}$ and $t_{AB}$. As described in Appendix 3.E, user $A$’s costs as a fraction of total costs work out to

$$f \equiv \frac{TC_A}{TC^o} = \frac{3 + z}{4} - \frac{1}{2} \sqrt{\frac{\alpha}{\alpha + \gamma} + z^2},$$

where $z \equiv \beta \cdot \gamma / ((\alpha + \gamma) \cdot (\beta + \gamma))$. The fraction $f$ depends on $\alpha$, $\beta$, and $\gamma$ only through the ratios $\beta/\alpha < 1$ and $\gamma/\alpha \leq 1$. It is readily shown that $f$ is a monotonically increasing function of $\beta/\alpha$ and $\gamma/\alpha$. At the upper limits with $\beta/\alpha \approx 1$ and $\gamma/\alpha = 1$, $f \approx 0.4535$. As $\beta/\alpha$ and $\gamma/\alpha$ approach their lower limits of 0, $f$ approaches $1/4$.

Figure 3.2: A PSNE with asymmetric costs: $\gamma \leq \alpha$, no queuing.

This example illustrates that although there exists a unique PSNE in terms of the aggregate departure rate, individual user departure rates can differ substantially and so can their costs. If the unit schedule delay cost parameters, $\beta$ and $\gamma$, are small compared to the cost of travel time, $\alpha$, one user’s fleet costs can be as little as one third the other user’s costs. This is because one user can concentrate its departures around $t^*$ without the other user wanting to reschedule its fleet because the gain from reducing schedule delay costs would be outweighed by the high costs of queuing delay. This suggests that equity of access to a bottleneck can be an issue with atomic users.
3.3.3 Extension to multiple users

The analysis in this section can be generalized to \( m > 2 \) users. Propositions 3.1 and 3.2 can be extended in a straightforward manner by following similar lines of reasoning. For example, condition (3.15) for existence of a PSNE without queuing, \( \lambda_{t,t'} \geq \gamma / (\alpha + \gamma) \), still holds. In the least restrictive case in which all \( m \) users depart at equal rates, this condition implies that \( 1/m \geq \gamma / (\alpha + \gamma) \), or \( (m - 1) \cdot \gamma \leq \alpha \). The intuition is similar to the case with \( m = 2 \). In a symmetric candidate PSNE, each user departs at rate \( s/m \) and contributes a fraction \( 1/m \) of the traffic. If user \( i \) reschedules one vehicle \( \Delta T \) units of time earlier, it reduces the vehicle’s late arrival cost by \( \gamma \cdot \Delta T \). The vehicle imposes a delay of \( 1/s \) on the \( (s/m) \cdot \Delta T \) of user \( i \)'s other vehicles that depart during the time interval. The unit cost of the delay is \( \alpha + \gamma \). User \( i \) benefits from rescheduling the vehicle unless \( \gamma \cdot \Delta T \leq (\alpha + \gamma) \cdot \Delta T/m \), which is equivalent to \( (m - 1) \cdot \gamma \leq \alpha \).

The nonexistence of a PSNE with queuing can also be established using the same approach for \( m > 2 \) as for \( m = 2 \). Candidate equilibrium departure rates for other users during early and late departures are still given by Eqs. (3.19) and (3.17) respectively. The candidate equilibrium is well-defined and unique, and it breaks down because any user can reduce its costs by rescheduling a portion of its fleet. Clearly, the condition \( (m - 1) \cdot \gamma \leq \alpha \) becomes more stringent as \( m \) increases so that a PSNE without queuing becomes progressively less plausible. Indeed, given any fixed values of \( \gamma > 0 \) and \( \alpha > 0 \), condition \( (m - 1) \cdot \gamma \leq \alpha \) necessarily fails if \( m \) is large enough. Thus, as \( m \to \infty \) the bottleneck model with atomic users does not converge in behavior to the non-atomic model which has a unique PSNE. The reason for this divergence is that even for large values of \( m \), users in the atomic model take into account how they affect queuing times. By contrast, in the standard bottleneck model users treat queuing times as given.

3.4 Nonuniqueness of equilibrium with heterogeneous atomic users

In this section we consider a modified version of the model in which users \( A \) and \( B \) differ in their preferred arrival times and can have fleets of different size. Arnott et al. (1987) studied this setting for the case of non-atomic users and we draw on some of their results. We show that, regardless of the relative size of parameters \( \alpha \) and \( \gamma \), a multiplicity of PSNE without queuing can exist with two atomic users if preferred arrival times differ within a certain range. The PSNE differ both in the timing of aggregate departures and total costs. Departures can begin earlier than, later than, or at the same time as in the social optimum. Thus, a PSNE can be inefficient even if there is no queuing.

Suppose user \( i, i = A, B \), has a fleet of \( N_i \) vehicles, each with a preferred arrival time of \( t_i^* \). Assume without loss of generality that \( t_B^* > t_A^* \). To begin, assume \( t_B^* \gg t_A^* \) so that the users can schedule their fleets at individually optimal time windows that do not overlap. Let \( t_{is} \) and \( t_{ie} \) denote the first and last departure times for user \( i \), and let superscript \( d \) denote the PSNE with disjoint arrival times. Each user departs at a rate \( s \) during the
3.4 Nonuniqueness of equilibrium with heterogeneous atomic users

interval that minimizes the total schedule delay costs of its fleet:

\[ (t_{As}^d, t_{Ae}^d) = \left( t_A^* - \frac{\gamma}{\beta + \gamma} \cdot \frac{N_A}{s}, t_A^* + \frac{\beta}{\beta + \gamma} \cdot \frac{N_A}{s} \right), \]  
\[ (t_{Bs}^d, t_{Be}^d) = \left( t_B^* - \frac{\gamma}{\beta + \gamma} \cdot \frac{N_B}{s}, t_B^* + \frac{\beta}{\beta + \gamma} \cdot \frac{N_B}{s} \right). \]  

(3.32)  

(3.33)

To rule out this uninteresting case we hereafter assume \( t_{Bs}^d < t_{Ae}^d \), which is equivalent to assuming that:

\[ t_B^* - t_A^* < \frac{\beta}{\beta + \gamma} \cdot \frac{N_A}{s} + \frac{\gamma}{\beta + \gamma} \cdot \frac{N_B}{s}. \]  

(3.34)

Define the auxiliary variable

\[ x = \frac{\beta}{\beta + \gamma} \cdot \frac{N_A}{s} + \frac{\gamma}{\beta + \gamma} \cdot \frac{N_B}{s} - (t_B^* - t_A^*) > 0. \]  

(3.35)

Variable \( x \) measures how much the users’ preferred departure schedules overlap, and thus the degree of “conflict” between them.

In addition to condition (3.34) we assume

\[ t_B^* - t_A^* > \left\| \frac{\gamma}{\beta + \gamma} \cdot \frac{N_B}{s} - \frac{\beta}{\beta + \gamma} \cdot \frac{N_A}{s} \right\|. \]  

(3.36)

Condition (3.36) assures that in the social optimum, derived next, some of user \( A \)’s fleet arrives late and some of user \( B \)’s fleet arrives early.

3.4.1 Social optimum

In the social optimum, vehicles depart at an aggregate rate of \( s \) over a connected time interval. In general, the optimal individual departure rates, \( r_A^*(t) \) and \( r_B^*(t) \), are not unique. However, given condition (3.36), there is a unique optimum in which all of user \( A \)’s fleet is dispatched before any of user \( B \)’s fleet. Some of user \( A \)’s fleet arrives late, and some of user \( B \)’s fleet arrives early. Let \( t_{As} \) be the time at which user \( A \) starts to depart. User \( A \) departs over the interval \((t_{As}, t_{As} + N_A/s)\), and user \( B \) departs over the interval \((t_{As} + N_A/s, t_{As} + (N_A + N_B)/s)\). Total costs are:

\[ TC = \frac{\beta \cdot s^2}{2} \cdot (t_A^* - t_{As})^2 + \frac{\gamma \cdot s}{2} \cdot \left( t_{As} + \frac{N_A}{s} - t_A^* \right)^2 \]  
\[ + \frac{\beta \cdot s^2}{2} \cdot \left( t_B^* - \left( t_{As} + \frac{N_A}{s} \right) \right)^2 + \frac{\gamma \cdot s}{2} \cdot \left( t_{As} + \frac{N_A + N_B}{s} - t_B^* \right)^2. \]  

(3.37)

The first-order condition for minimizing \( TC \) with respect to \( t_{As} \) yields:

\[ t_{As}^* = \frac{t_A^* + t_B^*}{2} - \frac{\gamma}{2 \cdot (\beta + \gamma)} \cdot \frac{N_A + N_B}{s} - \frac{1}{2} \cdot \frac{N_A}{s}. \]  

(3.38)
where superscript \( o \) again refers to the social optimum. Given condition (3.36), \( t_A < t_{Ae} < t_B^* \).

Substituting Eq. (3.38) into Eq. (3.37), total costs in the social optimum can be written as:

\[
TC^o = \frac{\delta}{2} \cdot \frac{N_A^2 + N_B^2}{s} + \frac{\beta + \gamma}{4} \cdot s \cdot x^2,
\]

where \( x \) is defined in Eq. (3.35). The first term in Eq. (3.39) equals total costs if condition (3.34) did not hold and the two users traveled in disjoint intervals. The second term in Eq. (3.39) is the additional costs incurred due to overlap in the users’ preferred arrival times. This term is an increasing, quadratic function of the degree of conflict, \( x \).

### 3.4.2 No-toll equilibrium with heterogeneous non-atomic users

Before considering the PSNE with users \( A \) and \( B \), we briefly discuss the analogous no-toll PSNE with heterogeneous non-atomic users. In this case there is a measure \( N_A \) of non-atomic users with a preferred arrival time \( t_A^* \), and a measure \( N_B \) of non-atomic users with a preferred arrival time \( t_B^* \). Arnott et al. (1987) show that when conditions (3.34) and (3.36) both hold, the equilibrium queuing pattern has two peaks. The first peak corresponds to on-time arrival for users with preferred arrival time \( t_A^* \), and the second peak to on-time arrival for users with preferred arrival time \( t_B^* \). Total costs are

\[
TC^n = \delta \cdot \frac{N_A^2 + N_B^2}{s} + \frac{\beta N_A + \gamma N_B}{2} \cdot x.
\]

The first term in (3.40) gives total costs if the equilibrium departure schedules of the two user groups do not conflict. This term is twice the first term in (3.39), just as total costs with homogeneous users in Eq. (3.7) are twice total costs in the social optimum in Eq. (3.11). The second term in (3.40) is a linear function of \( x \). This contrasts with the second term in (3.39) which is a quadratic function of \( x \). Thus, a small degree of conflict between the two user groups raises total costs disproportionately more in the no-toll equilibrium than the social optimum. In the next subsection we use the difference between total costs in the no-toll equilibrium and social optimum as a metric to assess the efficiency achieved from self-internalization of congestion by atomic users.

### 3.4.3 PSNE

A PSNE without queuing exists if \( t_{Ae} < t_B^* \) since the two users then do not arrive late at the same time and do not compete to complete their trips as soon as possible. Indeed, a continuum of PSNE exists. To see this, assume that both users depart in the period \([t_A^*, t_{Ae}]\) and capacity is fully used. Suppose user \( A \) advances a vehicle’s departure from \( t_2 \) to \( t_1 \), where \( t_A^* \leq t_1 < t_2 \leq t_{Ae} \). User \( A \)'s fleet costs change by:

\[
\Delta C_A = -C_A(t_2) + C_A(t_1) = -\gamma \cdot (t_2 - t_A^*) + \gamma \cdot (t_1 - t_A^*) + \frac{\alpha + \gamma}{s} \cdot \int_{t_1}^{t_2} r_A(u) du
\]

\[
= \left( (\alpha + \gamma) \cdot \Lambda_{t_1,t_2}^A - \gamma \right) \cdot (t_2 - t_1).
\]
3.4 Nonuniqueness of equilibrium with heterogeneous atomic users

User A does not benefit from the rescheduling if

\[ \lambda_{t_1,t_2}^A \geq \frac{\gamma}{\alpha + \gamma} . \]

This condition is satisfied for any choice of \( t_1 \) and \( t_2 \) as long as:

\[ r_A(t) \in \left[ \frac{\gamma}{\alpha + \gamma} \cdot s, s \right], \quad t \in [t_A^*, t_A^e] . \]  

(3.41)

User B’s departure rate is then \( r_B(t) = s - r_A(t) \in [0, \alpha \cdot s / (\alpha + \gamma)] \). There is no restriction on \( r_B(t) \) because User B arrives early throughout \([t_A^*, t_A^e]\). Thus, any departure profile satisfying Eq. (3.41) is consistent with a PSNE. However, different departure profiles generally result in different total costs.

An exhaustive treatment of all PSNE in this example would be tedious. Attention is limited to two illustrative cases. In Case 1, user A schedules its entire fleet during the same interval it would choose if user B did not exist. Hence, \( r_A(t) = s \) during the interval \((t_{As}, t_{Ae})\) given by Eq. (3.32). User B departs at rate \( s \) during the ensuing interval \((t_{Ae}, t_{Be})\). In Case 2, user A departs at rate \( s \) during the interval \((t_{As}, t_A^*)\). During the next interval \((t_A^*, t_{Ae})\), user A departs at the minimum rate consistent with Eq. (3.41), \( r_A(t) = \gamma \cdot s / (\alpha + \gamma) \). User B departs at the complementary rate \( r_B(t) = \alpha \cdot s / (\alpha + \gamma) \). During the final interval \((t_{Ae}, t_{Be})\), user B departs at rate \( s \). Cases 1 and 2 are now examined in turn.

3.4.3.1 PSNE for Case 1

Figure 3.3 depicts a PSNE conforming with Case 1. Note that, although there is no queueing in the PSNE, the value of parameter \( \alpha \) is relevant because it determines the maximum rate at which User B can depart during the central period.

Since all of user A’s fleet departs before any of user B’s fleet,

\[ t_{Ae}^1 - t_{As}^1 = \frac{N_A}{s} , \]  

(3.42)

where superscript 1 denotes Case 1. User A can schedule a vehicle either just before \( t_{As}^1 \) or just after \( t_{Ae}^1 \) without imposing a congestion delay on other vehicles in its fleet. The first and last vehicles must therefore incur equal schedule delay costs:

\[ \beta \cdot (t_{As}^1 - t_{Ae}^1) = \gamma \cdot (t_{Ae}^1 - t_A^*) . \]  

(3.43)

Together, Eqs. (3.42) and (3.43) imply:

\[ t_{As}^1 = t_A^* - \frac{\gamma}{\beta + \gamma} \cdot \frac{N_A}{s} , \]  

(3.44)

\[ t_{Ae}^1 = t_A^* + \frac{\beta}{\beta + \gamma} \cdot \frac{N_A}{s} . \]  

(3.45)

As noted above, the departure interval \((t_{As}^1, t_{Ae}^1)\) coincides with the interval \((t_{As}^d, t_{Ae}^d)\) given
by Eq. (3.32). User B departs in a connected time interval immediately after user A. Hence,

\[ t_{Bs}^1 = t_{Ae}^* = t_A^* + \frac{\beta}{\beta + \gamma} \cdot \frac{N_A}{s}, \quad (3.46) \]

\[ t_{Be}^1 = t_B^* + \frac{\beta}{\beta + \gamma} \cdot \frac{N_A}{s} + \frac{N_B}{s}. \quad (3.47) \]

A notable feature of the departure schedule given by Eqs. (3.44)-(3.47) is that it does not depend on \( t_B^* \). To be consistent with the assumption that some of user B’s fleet arrives early we require \( t_{Ae}^1 < t_B^* \), or:

\[ t_B^* - t_A^* > \frac{\beta}{\beta + \gamma} \cdot \frac{N_A}{s}. \quad (3.48) \]

Case 1 is a PSNE if neither user has an incentive to deviate from it. User A clearly has no incentive to deviate because its fleet costs are the minimum possible. Two conditions must be satisfied for user B. First, user B cannot gain by rescheduling its last vehicle from \( t_{Be}^1 \) to some time \( \hat{t} \) between \( t_{As}^1 \) and \( t_{Ae}^1 \). This condition is satisfied since rescheduling would impose queuing delay on all vehicles that depart after \( \hat{t} \) without reducing schedule delay costs for user B’s fleet.

Second, user B cannot gain by rescheduling its last vehicle from \( t_{Be}^1 \) to just before the beginning of the travel period at \( t_{As}^1 \). The condition is \( \beta \cdot (t_B^* - t_{As}^1) > \gamma \cdot (t_{Be}^1 - t_B^*), \) which reduces to:

\[ t_B^* - t_A^* \geq \frac{\gamma}{\beta + \gamma} \cdot \frac{N_B}{s}. \quad (3.49) \]
3.4 Nonuniqueness of equilibrium with heterogeneous atomic users

Conditions (3.48) and (3.49) together guarantee that condition (3.36) is satisfied. The timing of departures in Case 1 and the social optimum can be compared using Eqs. (3.44) and (3.38):

$$t_{As}^1 - t_{As}^o = \frac{1}{2} \cdot \left( \frac{\beta}{\beta + \gamma} \cdot \frac{N_A}{s} + \frac{\gamma}{\beta + \gamma} \cdot \frac{N_B}{s} - (t^*_{B} - t^*_{A}) \right) = \frac{x}{2}.$$

Since $x > 0$, $t_{As}^1 > t_{As}^o$: departures in Case 1 begin later than in the social optimum. To see why, note that while early and late arrivals are balanced optimally for user $A$, user $B$’s arrivals begin inefficiently late because user $A$ occupies the bottleneck until $t_{As}^1$ and effectively squeezes user $B$ out. As shown in Appendix 3.F, total costs in Case 1 can be written as:

$$TC^1 = \frac{\delta}{2} \cdot \frac{N_A^2 + N_B^2}{s} + \frac{\beta + \gamma}{2} \cdot s \cdot x^2 = TC^o + \frac{\beta + \gamma}{4} \cdot s \cdot x^2. \quad (3.50)$$

Total costs are an increasing quadratic function of the degree of conflict, $x$. This is similar to the social optimum (cf. Eq. (3.39)), but unlike the non-atomic equilibrium where the dependence is linear (cf. Eq. (3.40)). The last expression in Eq. (3.50), $(\beta + \gamma) sx^2/4$, measures inefficiency of the PSNE in Case 1. This inefficiency can be expressed as a price of anarchy, $PA$, using the ratio

$$PA = \frac{TC^1 - TC^o}{TC^o}.$$

$PA$ is bounded below by 0. As shown in Appendix 3.F, it is bounded above by $1/5$. Thus, despite the fact that there is no queuing in the PSNE for Case 1, total costs can exceed the socially optimal level by up to 20 percent.

Efficiency of the PSNE can also be measured relative to the non-atomic equilibrium using the index

$$w \equiv \frac{TC^o - TC^1}{TC^o - TC^o}. \quad (3.51)$$

The numerator of (3.51) is the reduction in total costs achieved by self-internalization of congestion by the atomic users. The denominator is the reduction in total costs realized at the social optimum. Index $w$ reaches its maximum value of 1 when $x = 0$. Appendix 3.F establishes that it is bounded below by $6/7$. Self-internalization of congestion by the atomic users independently therefore achieves most of the potential benefits from coordinating departure times centrally.

3.4.3.2 PSNE for Case 2

Figure 3.4 depicts an example of a PSNE in Case 2 using the same parameter values as for Figure 3.3. During the initial interval, $(t_{As}^2, t_{Ae}^2)$, user $A$ departs at rate $s$. User $B$ begins to depart immediately after $t_{As}^2$, and during the interval $(t_{Ae}^2, t_{Be}^2)$ the two users depart simultaneously with an aggregate rate of $s$. User $A$ departs at rate $r_A(t) = \gamma \cdot s / (\alpha + \gamma)$, and user $B$ departs at rate $r_B(t) = \alpha \cdot s / (\alpha + \gamma)$. During the final interval, $(t_{Ae}^2, t_{Be}^2)$, user $B$ departs at rate $s$. 

65
Chapter 3 Bottleneck model with atomic users

The PSNE for Case 2 is solved in Appendix 3.F. The three transition times work out to:

\[ t^2_{As} = t^*_A - \frac{\alpha + \gamma}{\alpha + \beta + \gamma} \cdot \frac{N_A}{s}, \]  
(3.52)

\[ t^2_{Ae} = t^2_{Bs} = t^*_A + \frac{\beta \cdot (\alpha + \gamma)}{\gamma (\alpha + \beta + \gamma)} \cdot \frac{N_A}{s}, \]  
(3.53)

\[ t^2_{Be} = t^*_A + \frac{\beta}{\alpha + \beta + \gamma} \cdot \frac{N_A}{s} + \frac{N_B}{s}. \]  
(3.54)

Similar to Case 1, the timing of the PSNE given by Eqs. (3.52)-(3.54) does not depend on \( t^*_B \). As shown in Appendix 3.F, departures in the PSNE of Case 2 can begin either earlier or later than in the social optimum. As \( t^*_B \) decreases, the social optimum begins earlier (cf. Eq. (3.38)), whereas the PSNE does not (cf. Eq. (3.52)). Hence, as user B’s preferred schedule moves closer to user A’s and the conflict between them grows, the PSNE in Case 2 shifts from beginning too early to beginning too late.

Unlike for Case 1, there is no simple formula for total costs in Case 2. Also unlike Case 1, total costs do not reach the socially optimal minimum value at the point where \( t^2_{As} \) coincides with \( t^*_A \). Thus, the price of anarchy is always positive. This is because both users depart during the interval \((t^*_A, t^2_{Ae})\), with user A arriving late and user B arriving early. To see that simultaneous departure is inefficient, consider a vehicle of user B arriving at \( t_1 \) and a vehicle of user A arriving at \( t_2 > t_1 \). If the two vehicles exchanged slots, user A’s vehicle would arrive less late and user B’s vehicle would arrive less early. The two users might agree to such an exchange if they cooperated, but the switch cannot occur in a Nash equilibrium.

Figure 3.4: PSNE for Case 2 with \( t^*_B > t^*_A \). \( N_A = 1 \), \( N_B = 1 \), \( s = 1 \), \( \alpha = 2 \), \( \beta = 1 \), \( \gamma = 2 \), \( t^*_A = 0 \), and \( t^*_B = 2/3 \).
We conclude this section with a numerical example featuring the same parameter values as those used to construct Figures 3.3 and 3.4: \( N_A = 1, N_B = 1, s = 1, \alpha = 2, \beta = 1, \gamma = 2, t^*_A = 0, \) and \( t^*_B = 2/3. \) Departures and arrivals last for two hours. The social optimum begins at \( t^o_A = -5/6, \) and total costs are \( TC^o = 3/4. \) The PSNE for Case 1 begins at \( t^1_A = -2/3, \) and total costs are \( TC^1 = 5/6 \) which is 11.1\% higher than \( TC^o. \) User \( A's \) total costs are 1/3, and user \( B's \) total costs are 1/2. User \( B's \) total costs are 50\% higher than user \( A's \) total costs even though the two users have the same size of fleet and the same unit costs \( \alpha, \beta, \) and \( \gamma. \) The PSNE for Case 2 begins at \( t^2_A = -4/5, \) and total costs are \( TC^2 = 0.81 \) which is 8.4\% higher than \( TC^o \) but 2.4\% lower than \( TC^1. \) User \( A's \) total costs are 2/5, and user \( B's \) total costs are 0.413. User \( B's \) total costs are just 3.3\% higher than user \( A's \) total costs so that the PSNE in Case 2 is both more efficient and more equitable than in Case 1. Finally, note that if user \( B's \) desired arrival time is increased from \( t^*_B = 2/3 \) to \( t^*_B = 4/5, t^*_A \) increases to \(-23/30\) while \( t^2_A \) remains at \(-4/5. \) The PSNE in Case 2 then begins later than the social optimum.

3.5 Conclusions

In this chapter we have explored the existence and uniqueness of pure strategy Nash equilibrium (PSNE) in the bottleneck model with atomic users. We consider a simple case featuring two users with piecewise linear and independent trip-timing preferences for each vehicle in their fleets. We show that if users are identical, and \( \gamma > \alpha, \) a PSNE does not exist. By contrast, if \( \gamma \leq \alpha, \) multiple PSNE do exist in which no queuing occurs. The aggregate departure profile is unique. However, there exists a continuum of departure-time schedules for the two users that differ in the share of total costs borne by each user. We also consider a setting in which the two users differ in their preferred arrival times, \( t^*_A \) and \( t^*_B. \) For a range of values of \( t^*_B - t^*_A, \) there exists a continuum of PSNE that differ in the timing of trips, total costs, and the relative burden of costs borne by each user.

The potential nonexistence of PSNE with atomic users is, we believe, the most significant of the results. In the standard bottleneck model a PSNE exists under relative general assumptions. It fails to exist only if trip-timing preferences have discontinuities such as discrete penalties for late arrival or discontinuities in the cost of travel time as a function of duration (Lindsey, 2004). In the model with non-atomic users nonexistence is a problem even with smooth preferences because atomic users face conflicting incentives. On the one hand they prefer to spread out departures of their fleets in order to avoid self-imposed queuing delays. If there is only one user this results in a socially-optimal departure schedule with no queuing. But when other users are present, an atomic user has an incentive to force vehicles into the departure stream near the peak in order to reduce its fleet’s schedule delay costs. A user’s best response can be to preempt other users by scheduling a mass departure near \( t^* \) and hogging bottleneck capacity. Yet mass departures are inconsistent with PSNE, as we have shown.

The out-of-equilibrium behaviour in the bottleneck model bears some resemblance to firm behaviour in one-dimensional (Hotelling) spatial competition models (see Anderson et al. (1992), chapter 8). In location-choice games with parametric prices firms leapfrog each other to gain market share. The outcome depends on the number of firms. With
two or four firms a unique PSNE exists. But with five firms there are multiple PSNE, and
with three firms a PSNE does not exist. In simultaneous location- and price-choice games
with two firms a PSNE does not exist. Firms face conflicting location-choice incentives
that are broadly similar to the incentives of atomic users picking departure times in the
bottleneck model. Moving closer to the centre (analogous to departing close to \( t^* \)) gains
firms market share, but moving away softens price competition.

Preliminary analysis suggests that the difference between the standard bottleneck model
and the model with atomic users remains as the number of atomic users increases because
each user still recognizes how its scheduling choices affect queuing delays. Thus, simple
congestion-prone systems with atomic users can exhibit fundamentally different dynamic
behavior than with non-atomic users. Dynamic systems with non-atomic users have been
well studied in the literature using the bottleneck model as well as various flow-congestion
models. Yet, to our knowledge, this chapter is the first to examine the existence and
uniqueness of equilibrium in dynamic systems with atomic users that make scheduling
decisions simultaneously.

Our analysis is exploratory and can be extended in various directions. One obvious
priority is to determine how the bottleneck model can be modified to restore existence
of equilibrium without restricting parameter values. Possibilities include: extension of
the solution concept to mixed strategies or sequential decision-making (e.g., Stackelberg
equilibrium), and a more general specification of scheduling preferences based on activity
analysis.

Another direction for future research is to investigate more fully the benefits of coor-
dination or other forms of cooperation. Self-internalization of congestion externalities by
an atomic agent is one form of coordination, and its potential for welfare gains is ap-
parent from the examples in Section 3.4. Cooperation between atomic agents can also
be beneficial as shown by the examples in which the price of anarchy is positive. Major
freight shippers might cooperate by harmonizing their delivery schedules although doing
so without triggering concerns about illegal collusion might be tricky. Governments
can also implement or encourage cooperative behavior. One example are staggered work
hours and flextime programs that entail coordination on target arrival times (i.e., \( t^* \)).
Staggered work hours were introduced in the 1970s in Washington, D.C., and Ottawa,
Canada. Gutiérrez-i-Puigarnau and Van Ommeren (2012) report that one in three Ger-
mans firms uses staggered working hours. Mun and Yonekawa (2006) note that in Japan,
in 1998, 8% of employees worked at a firm with flexible working hours, and in the US,
in 1994-1997, less than 6% of employees had a formal flexible working arrangement, but
28% of all full-time workers varied their working times to some degree.

A recent example of a government-sponsored program is the Dutch Spitsmijden exper-
iment in which drivers are offered rewards for avoiding peak-hour travel by retiming their
trips, choosing another mode, or teleworking (Ben-Elia and Ettema, 2009, 2011). The
Spitsmijden experiment differs from other travel-demand measures such as road pricing
in that participation is voluntary. Large employers in the test regions have shown interest
in participating as a way to curb congestion and to attract or retain employees (Ben-Elia
and Ettema, 2009). However, it is an open question whether behavioral changes induced
by a reward system will persist after a program is terminated (Ben-Elia and Ettema,
2011).
Appendix 3.A Glossary of notation

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>c(t)</td>
<td>Generalized cost of a trip that departs at t</td>
</tr>
<tr>
<td>C_i(t)</td>
<td>Marginal private cost to user i of scheduling a vehicle at time t</td>
</tr>
<tr>
<td>C_i^+(t)</td>
<td>Cost for user i of adding a vehicle to the departure schedule at time t</td>
</tr>
<tr>
<td>C_i^-(t)</td>
<td>Cost for user i of removing a vehicle from the departure schedule at time t</td>
</tr>
<tr>
<td>N</td>
<td>Aggregate number of vehicles</td>
</tr>
<tr>
<td>N_i</td>
<td>Fleet size of user i</td>
</tr>
<tr>
<td>Q(t)</td>
<td>Length of the queue at time t</td>
</tr>
<tr>
<td>Q̇</td>
<td>Rate of change of the queue length</td>
</tr>
<tr>
<td>r(t)</td>
<td>Aggregate departure rate from the origin at time t</td>
</tr>
<tr>
<td>r_i(t)</td>
<td>Departure rate of user i from the origin at time t</td>
</tr>
<tr>
<td>s</td>
<td>Capacity of the bottleneck</td>
</tr>
<tr>
<td>t</td>
<td>Time of departure</td>
</tr>
<tr>
<td>t_a</td>
<td>Time of arrival</td>
</tr>
<tr>
<td>t_e</td>
<td>Last moment of operation (moment of the last arrival)</td>
</tr>
<tr>
<td>t_s</td>
<td>First moment of operation (moment of the first arrival)</td>
</tr>
<tr>
<td>t^*</td>
<td>Desired arrival time</td>
</tr>
<tr>
<td>t_i^*</td>
<td>Desired arrival time of user i</td>
</tr>
<tr>
<td>t̂</td>
<td>Time for which a user arrives on time</td>
</tr>
<tr>
<td>T(t)</td>
<td>Travel delay of a vehicle arriving at time t</td>
</tr>
<tr>
<td>α</td>
<td>Value of travel time for</td>
</tr>
<tr>
<td>β</td>
<td>Value of early schedule delay</td>
</tr>
<tr>
<td>γ</td>
<td>Value of late schedule delay</td>
</tr>
</tbody>
</table>

Table 3.1: Glossary of notation.

Appendix 3.B Proof of Lemma 3.2

Suppose that user A schedules a mass departure at time t. Clearly, user B will not depart either at t or immediately after t since it would be better to depart either just before the mass or some time after t when the queue caused by the mass departure has diminished or disappeared. To complete the proof we will show that user A will never schedule a mass departure at t unless user B is scheduling vehicles either at t or immediately after t.

Thus, suppose that user A is considering a mass departure of M vehicles at time t, and assume that user B does not depart during the period [t, t̅], where t̅ > t. Variables M, t, and t̅ are held fixed throughout the proof. The logic of the proof depends on whether or not a queue exists at t.
3.B.1 A queue exists at \( t \)

Assume \( Q(t) > 0 \). Pick any time \( t' \in (t, \bar{t}] \) such that \( Q(t') > 0 \). We will show that user \( A \) is better off postponing the mass departure from \( t \) to \( t' \). Let \( m \in [0, M] \) index vehicles in the order they are positioned in the mass, and let \( c(m, t) \) denote the cost incurred by vehicle \( m \) when the mass departs at \( t \):

\[
c(m, t) = \alpha \cdot \left( T(t) + \frac{m}{s} \right) + \left\{ \begin{array}{ll}
\beta \cdot \left( t^* - t - T(t) - \frac{m}{s} \right) & \text{if } t + T(t) + \frac{m}{s} \leq t^* \\
\gamma \cdot \left( t + T(t) + \frac{m}{s} - t^* \right) & \text{if } t + T(t) + \frac{m}{s} \geq t^*
\end{array} \right.
\]

If the mass is rescheduled to depart at \( t' \) instead, vehicle \( m \) incurs a cost:

\[
c(m, t') = \alpha \cdot \left( T(t') + \frac{m}{s} \right) + \left\{ \begin{array}{ll}
\beta \cdot \left( t^* - t' - T(t') - \frac{m}{s} \right) & \text{if } t' + T(t') + \frac{m}{s} \leq t^* \\
\gamma \cdot \left( t' + T(t') + \frac{m}{s} - t^* \right) & \text{if } t' + T(t') + \frac{m}{s} \geq t^*
\end{array} \right.
\]

Now, vehicle \( m \) arrives at the same time whether the mass is scheduled at \( t \) or \( t' \) since a queue remains at \( t' \) and user \( B \) does not depart between \( t \) and \( t' \). Hence \( t' + T(t') = t + T(t) \), and \( c(m, t') \) can be written:

\[
c(m, t') = \alpha \cdot \left( T(t) + \frac{m}{s} \right) - \alpha \cdot (t' - t) + \left\{ \begin{array}{ll}
\beta \cdot \left( t^* - t - T(t) - \frac{m}{s} \right) & \text{if } t + T(t) + \frac{m}{s} \leq t^* \\
\gamma \cdot \left( t + T(t) + \frac{m}{s} - t^* \right) & \text{if } t + T(t) + \frac{m}{s} \geq t^*
\end{array} \right.
\]

\[
= c(m, t) - \alpha \cdot (t' - t) < c(m, t).
\]

Postponing the mass departure therefore reduces costs for all \( M \) vehicles in the mass.

3.B.2 No queue exists at \( t \)

Assume \( Q(t) = 0 \) and set \( t' = \bar{t} \). If \( M \leq s \cdot (t' - t) \), vehicles provisionally scheduled in the mass can instead be rescheduled to depart at rate \( s \) for a period of duration \( M/s \). The \( M \) vehicles arrive at the same time as with the mass departure, but without incurring a queuing delay. If \( M > s \cdot (t' - t) \), then \( s \cdot (t' - t) \) of the vehicles can be rescheduled to depart at rate \( s \) for a period \( t' - t \), and the remaining \( M - s \cdot (t' - t) \) vehicles can be rescheduled to depart in a mass at \( t' \). All \( M \) vehicles will arrive at the same time as they do in the mass departure at \( t \). The first \( s \cdot (t' - t) \) vehicles do not queue at all, and the last \( M - s \cdot (t' - t) \) vehicles do not queue during \([t, t']\). Total queuing costs are therefore reduced, and total schedule delay costs are unchanged. Postponing the mass departure again reduces fleet costs, so that the mass departure is not optimal. QED

Appendix 3.C Proof of Lemma 3.4

It suffices to show that user \( A \)'s vehicles that depart after \( t_q \) incur lower total costs in the revised schedule than in the candidate PSNE schedule.
3.C.1 Costs in the candidate equilibrium

Total queuing time in the candidate PSNE is measured by area $ABC$ in Figure 3.1. User $A$ incurs half of this delay. User $A$’s queuing time costs are therefore $\alpha \cdot (t^* - \hat{t}) \cdot s \cdot (t_e - t_q) / 4$. The number of user $A$’s vehicles that queue and arrive early is $s \cdot (t_e - t_q) / 2$, and their average schedule delay cost is $\beta \cdot (t^* - t_q) / 2$. The number of vehicles that queue and arrive late is $s \cdot (t_e - t^*) / 2$, and their average schedule delay cost is $\gamma \cdot (t_e - t^*) / 2$. User $A$’s queuing vehicles therefore incur total costs of

$$TC^e = \frac{\alpha \cdot s}{4} \cdot (t^* - \hat{t}) \cdot (t_e - t_q) + \frac{\beta \cdot s}{4} \cdot (t^* - t_q)^2 + \frac{\gamma \cdot s}{4} \cdot (t_e - t^*)^2,$$

where superscript $e$ denotes the candidate PSNE. Substituting Eqs. (3.28), (3.29), and (3.30) into Eq. (3.55) gives

$$TC^e = \frac{\beta^2 \cdot (\alpha^2 + \alpha \cdot (\gamma - \beta) + 3 \cdot \beta \cdot \gamma)}{8 \cdot s \cdot (\alpha + \beta)^2 \cdot (\beta + \gamma)} \cdot N^2.$$  \hspace{1cm} (3.56)

3.C.2 Costs in the revised schedule

Time $\hat{t}$ in Figure 3.1 is defined by the condition

$$\frac{\alpha}{\alpha - \beta} \cdot s \cdot (\hat{t} - t_q) + \frac{\alpha}{\alpha + \gamma} \cdot s \cdot (\hat{t} - \hat{t}) = s \cdot (\hat{t} - t_q).$$

This solves to yield

$$\hat{t} = t^* - \frac{\beta \cdot (\gamma - \alpha) \cdot (\gamma - \beta)}{2 \cdot \gamma \cdot (\alpha + \beta) \cdot (\beta + \gamma)} \cdot \frac{N}{s} < t^*,$$ \hspace{1cm} (3.57)

where the inequality follows from $\gamma > \alpha > \beta$.

The number of user $A$’s vehicles that depart after $\hat{t}$ and arrive early in the revised schedule is $\gamma \cdot s \cdot (t^* - \hat{t}) / (\alpha + \gamma)$. These vehicles incur an average schedule delay cost of $\beta \cdot (t^* - \hat{t}) / 2$. The number of user $A$’s vehicles that depart after $\hat{t}$ and arrive late is $\gamma \cdot s \cdot (t_e - t^*) / (\alpha + \gamma)$ and their average schedule delay cost is $\gamma \cdot (t_e - t^*) / 2$. Total costs are therefore

$$TC^d = \frac{\beta \cdot \gamma \cdot s}{2 \cdot (\alpha + \gamma)} \cdot (t^* - \hat{t})^2 + \frac{\gamma^2 \cdot s}{2 \cdot (\alpha + \gamma)} \cdot (t_e - t^*)^2,$$ \hspace{1cm} (3.58)

where superscript $d$ denotes the deviant schedule. Substituting Eqs. (3.57) and (3.28) into Eq. (3.58) gives

$$TC^d = \frac{\alpha^2 \cdot (\beta^2 - 3 \cdot \beta \cdot \gamma + 4 \cdot \gamma^2) + 2 \cdot \alpha \cdot \beta \cdot \gamma \cdot (3 \cdot \gamma - \beta) + \beta \cdot \gamma^2 \cdot (\beta + \gamma)}{8 \cdot \gamma \cdot (\alpha + \beta)^2 \cdot (\alpha + \gamma) \cdot (\beta + \gamma) \cdot s} \cdot N^2.$$ \hspace{1cm} (3.59)

Given Eqs. (3.56) and (3.59), the costs saved by deviating from the candidate PSNE
work out to

\[ TC^e - TC^d = \frac{\beta^2 \cdot (\alpha - \gamma)^2 \cdot (\alpha \cdot \gamma - \beta^2 + 2 \cdot \beta \cdot \gamma)}{8 \cdot s \cdot \gamma \cdot (\alpha + \beta)^2 \cdot (\alpha + \gamma) \cdot (\beta + \gamma)} \cdot N^2 > 0. \]

Since the cost saving is positive, the candidate PSNE is not a PSNE. QED

Appendix 3.D  Proposition 3.2: Proof of Result 4 for mass departures

Suppose that user A deviates from the candidate PSNE by scheduling multiple mass departures. We prove that such a deviation cannot reduce user A’s costs. The proof is done in two steps. We first show that any deviation with multiple mass departures does not achieve lower costs than a deviation with a single mass departure launched before \( t^* \) (3.D.1). We then show that this deviation is not gainful if the bounds on departure rates stated in Proposition 3.2 are satisfied (3.D.2).

3.D.1 Optimality of the single mass departure deviation

We show that any deviation from the PSNE involving multiple mass departures is dominated by a single mass departure. The proof involves establishing the following three results for either user: (i) Fleet costs can be (weakly) reduced by rescheduling any vehicles that suffer queuing delay, but are not part of a mass, to a period without queuing. (ii) Fleet costs can be (weakly) reduced by rescheduling any vehicles in a mass departure after \( t^* \) to a period without queuing. (iii) Any deviation with multiple mass departures launched before \( t^* \) entails strictly higher fleet costs than a deviation with a single mass departure. These three results establish that the candidate PSNE need only be tested against a single mass departure launched before \( t^* \).

Result i: When a queue exists, user A is willing to depart at a positive and finite rate only if condition (3.20) is satisfied. For early arrivals this requires \( r_B(t) = \alpha \cdot s / (\alpha - \beta) > s \). Since \( r_B(t) < s \) in the candidate PSNE, condition (3.20) is violated, and user A will not deviate by departing early when there is a queue. For late arrivals, user A is willing to depart at a positive and finite rate when there is a queue only if \( r_B(t) = \alpha \cdot s / (\alpha + \gamma) \). Proposition 3.2 stipulates that \( r_B(t) \leq \alpha \cdot s / (\alpha + \gamma) \). If \( r_B(t) < \alpha \cdot s / (\alpha + \gamma) \), user A is better off scheduling its vehicles later. If \( r_B(t) = \alpha \cdot s / (\alpha + \gamma) \), user A is indifferent between departing or not, and would not gain by rescheduling vehicles late when there is a queue.

Result ii: Assume that the last mass departure is launched at a time of late arrivals. We show that rescheduling vehicles in the mass to a later period in which they do not incur queuing delay is beneficial. By induction, it then follows that all mass departures launched at times of late arrivals can be gainfully rescheduled.

Suppose the last mass departure is launched at time \( t_L \) and comprises \( M \) vehicles. Assume first that there is a queue at \( t_L \). We show that postponing the mass departure to the moment when the queue disappears (weakly) reduces fleet costs. Let \( m \in [0, M] \)
index vehicles in the order they are positioned in the mass, and let \( c(m, t_L) \) denote the cost incurred by vehicle \( m \):

\[
c(m, t_L) = \alpha \cdot \left( T(t_L) + \frac{m}{s} \right) + \gamma \cdot \left( t_L + T(t_L) + \frac{m}{s} - t^* \right).
\]

Now suppose the mass is postponed to \( t'_L > t_L \) when a queue still exists. Vehicle \( m \) now incurs a cost of:

\[
c(m, t'_L) = \alpha \cdot \left( T(t'_L) + \frac{m}{s} \right) + \gamma \cdot \left( t'_L + T(t'_L) + \frac{m}{s} - t^* \right). \tag{3.60}
\]

Since only user \( B \) departs during \( (t_L, t'_L) \),

\[
T(t'_L) = T(t_L) + \int_{t_L}^{t'_L} \frac{r_B(u) - s}{s} du = T(t_L) - (t'_L - t_L) \cdot \left( 1 - \lambda_{t_L, t'_L}^B \right). \tag{3.61}
\]

Substituting (3.61) in (3.60), we get:

\[
c(m, t'_L) = c(m, t_L) - \alpha \cdot (t'_L - t_L) \cdot \left( 1 - \lambda_{t_L, t'_L}^B \right) + \gamma \cdot (t'_L - t_L) \cdot \lambda_{t_L, t'_L}^B.
\]

Postponing the mass departure changes costs by:

\[
c(m, t'_L) - c(m, t_L) = (t'_L - t_L) \cdot \left[ \lambda_{t_L, t'_L}^B \cdot (\alpha + \gamma) - \alpha \right] \leq 0,
\]

where the inequality follows from the condition \( r_B(t) \leq \alpha \cdot s / (\alpha + \gamma) \) in Proposition 3.2.

We conclude that if there is a queue when the last mass departs, the mass departure can be postponed to the time where the queue just disappears without increasing costs (later vehicles are not affected by postponing the mass).

Now assume there is no queue at \( t_L \) when the last mass is launched. We show that vehicles in the mass can be rescheduled later to a time when they do not face a queue, and that doing so does not increase user \( A \)'s fleet costs. Consider the cost of the last (i.e., \( M \)-th) vehicle in the mass:

\[
c(M, t_L) = \gamma \cdot (t_L - t^*) + (\alpha + \gamma) \cdot \frac{M}{s}.
\]

Because user \( A \) does not depart until the queue has disappeared (result i), the queue produced by the mass departure disappears at time \( t'_L \) where:

\[
\frac{M}{s} + \int_{t_L}^{t'_L} \frac{r_B(u) - s}{s} du = \frac{M}{s} - (t'_L - t_L) \cdot \left( 1 - \lambda_{t_L, t'_L}^B \right) = 0. \tag{3.62}
\]

A vehicle that departs at \( t'_L \) incurs a cost:

\[
c(t'_L) = \gamma \cdot (t'_L - t^*) = \gamma \cdot (t'_L - t_L) + \gamma \cdot (t_L - t^*) = \gamma \cdot (t_L - t^*) + \gamma \cdot \frac{M}{s} \cdot \left( 1 - \lambda_{t_L, t'_L}^B \right).
\]
Chapter 3 Bottleneck model with atomic users

where the last equality uses Eq. (3.62). As removing a vehicle from the mass opens up a (queue-free) slot at \( t'_L \), rescheduling the last vehicle to \( t'_L \) yields a change in cost of:

\[
c(t'_L) - c(M, t_L) = \frac{M}{s} \cdot \left[ \frac{\gamma}{1 - \lambda^B_{t_L,t'_L}} - (\alpha + \gamma) \right] \leq 0 ,
\]

where the inequality follows from condition \( r_B(t) \leq \alpha \cdot s/((\alpha + \gamma)) \) in Proposition 3.2.

This shows that, in any deviation from the candidate PSNE, the last mass departure launched at a time of late arrivals can be eliminated without increasing fleet costs. By induction, any mass departure launched at a time of late arrivals can be rescheduled without increasing fleet costs.

Next, we show that any deviation entailing multiple mass departures at times of early arrivals is dominated by scheduling a single mass departure before \( t^* \).

**Result iii:** Suppose that more than one mass departure is scheduled before \( t^* \). Assume the first mass is launched at time \( t_E \) with \( M \) vehicles, and the second mass is launched at time \( t'_E > t_E \). If the queue from the first mass disappears before \( t'_E \), fleet costs can be reduced by rescheduling vehicles in the first mass to depart at a rate equal to residual capacity \((s - r_B(t))\). Since user \( A \) does not depart in the original deviation until the queue has dissipated, the rescheduled vehicles in the alternative deviation escape queuing and arrive less early – thereby reducing both their queuing and schedule delay costs. If the queue from the first mass does not disappear before \( t'_E \), user \( A \) can still reduce its fleet costs by rescheduling \( s \cdot (t'_E - t_E) \cdot (1 - \lambda^B_{t_E,t'_E}) \) vehicles at a rate \( s - r_B(t) \) during \((t_E, t'_E)\), and letting the remaining \( M - s \cdot (t'_E - t_E) \cdot (1 - \lambda^B_{t_E,t'_E}) \) vehicles join the head of the second mass at \( t'_E \). The first set of vehicles rescheduled at residual capacity avoid queuing and incur lower early arrival costs because they arrive closer to \( t^* \). The second set of vehicles also incur lower schedule delay costs since they arrive later. They also incur lower queuing costs as well since they no longer queue between \( t_E \) and \( t'_E \). Vehicles in the original mass that departs at \( t'_E \) still depart and arrive at the same time because the same number of vehicles depart before them and the bottleneck operates at capacity throughout.

By induction, all but one of the mass departures launched before \( t^* \) can be eliminated in a way that decreases fleet costs. Thus, results (i)–(iii) show that a deviation with a single mass departure launched before \( t^* \) is the most viable deviation, of deviations entailing mass departures, from the candidate PSNE. In the following section we show that the cost-minimizing deviation does not reduce fleet costs with respect to the candidate PSNE.

### 3.D.2 The optimal mass departure

Following the logic of Appendix C.1, suppose user \( A \) launches a single mass departure of \( M \) vehicles at \( t_m < t^* \). User \( A \) will not include in the mass vehicles that were scheduled to depart before \( t_m \) because this would create a queue that lasts until after \( t_e \), and the cost of the delay imposed on \( A \)'s later vehicles would outweigh any benefit. User \( A \) will include in the mass all vehicles scheduled to depart after \( t_m \) that would be delayed by the mass. However, as discussed above, this would be counterproductive if all these vehicles were scheduled to depart before \( t^* \) because they would suffer not only queuing delay but
also greater early-arrival costs (the queue would disappear before $t^*$).

The mass departure is potentially beneficial only if it includes vehicles scheduled to depart both early and late. Therefore, the mass must be launched at a time $t_m < t^*$, and the queue must persist until a time $t_M > t^*$, where $t_M$ is defined by the condition:

$$\int_{t_m}^{t_M} r_A(u) \, du = M,$$

where $r_A(\cdot)$ is user $A$'s departure schedule in the candidate PSNE. Since the queuing costs incurred by the mass do not depend on when it is launched, $t_m$ should be chosen to minimize total schedule delay costs for vehicles in the mass. The first and last vehicles should therefore incur the same schedule delay cost:

$$\beta \cdot (t^* - t_m) = \gamma \cdot \left( t_m + \frac{M}{s} - t^* \right).$$

This implies

$$t_m = t^* - \frac{\gamma}{\beta + \gamma} \cdot \frac{M}{s}.$$ 

Vehicles in the mass incur total queuing time costs of $\alpha \cdot M^2 / (2 \cdot s)$, and total schedule delay costs of $\delta \cdot M^2 / (2 \cdot s)$. Total costs for the mass are therefore:

$$TC_m = \frac{\alpha + \delta}{2} \cdot \frac{M^2}{s}.$$

In the candidate PSNE, where $r_A(t) + r_B(t) = s$ for $t \in (t_m, t_M)$, the $M$ vehicles incur total costs of

$$TC_e = \beta \cdot \int_{t_m}^{t^*} r_A(u) \cdot (t^* - u) \, du + \gamma \cdot \int_{t^*}^{t_M} r_A(u) \cdot (u - t^*) \, du.$$

To complete the proof we must show that $TC_e \leq TC_m$. User $A$'s departure rate over the period $(t_m, t_M)$ must be consistent with condition (3.63). As shown in proving Result 3 of Proposition 3.2, during late arrivals $r_A(u)$ is bounded below by $\gamma \cdot s / (\alpha + \gamma)$. Let $r_E$ denote the minimum departure rate of user $A$ prior to $t^*$. It follows by straightforward algebra that

$$TC_e \leq M^2 \cdot \frac{(\beta \cdot \gamma^2 \cdot r_E \cdot s + (\alpha + \gamma) \cdot ((\beta + \gamma)^2 \cdot s^2 + r_E^2 \cdot \gamma^2 - 2 \cdot (\beta + \gamma) \cdot \gamma \cdot s \cdot r_E))}{2 \cdot (\beta + \gamma)^2 \cdot s^3}.$$

Setting the right-hand side of inequality (3.65) equal to Eq. (3.64) one obtains

$$r_E = \frac{s}{2} \cdot \frac{2 \cdot \alpha \cdot (\beta + \gamma) + \gamma \cdot (\beta + 2 \cdot \gamma) - \sqrt{\beta^2 \cdot \gamma^2 + 4 \cdot \alpha \cdot (\beta + \gamma)^2 \cdot (\alpha + \gamma)}}{\gamma \cdot (\alpha + \gamma)}.$$

If user $A$'s departure rate during early arrivals is at least $r_E$, departing in a mass can-
not reduce its fleet costs. The candidate PSNE is therefore robust to mass departure deviations.

Straightforward algebra leads to:

\[
0 < r_E < \frac{\gamma}{\alpha + \gamma} \cdot s,
\]

which shows that in the limit \( \gamma \to 0, r_E \to 0 \). QED

**Appendix 3.E** Asymmetric equilibrium costs with \( \gamma \leq \alpha \)

User A’s fleet costs in the PSNE depicted in Figure 3.2 are

\[
TC_A = r_E \cdot (t_{BA} - t^*_A) \cdot \beta \cdot \left( t^* - \frac{t^*_s + t_{BA}}{2} \right) + (s - r_E) \cdot (t^* - t_{BA}) \cdot \beta \cdot \frac{t^* - t_{BA}}{2} + \frac{\alpha}{\alpha + \gamma} \cdot s \cdot (t_{AB} - t^*) \cdot \gamma \cdot \frac{t_{AB} - t^*}{2} + \frac{\gamma}{\alpha + \gamma} \cdot s \cdot (t_e - t_{AB}) \cdot \gamma \cdot \left( \frac{t_{AB} + t^*_s}{2} - t^* \right)
\]

\[
= \frac{\beta \cdot r_E}{2} \cdot \left( (t^* - t^*_s)^2 - (t^* - t_{BA})^2 \right) + \frac{\beta}{2} \cdot (s - r_E) \cdot (t^* - t_{BA})^2 + \frac{\alpha \cdot \gamma \cdot s}{2 \cdot (\alpha + \gamma)} \cdot (t_{AB} - t^*)^2 + \frac{\gamma^2 \cdot s}{2 \cdot (\alpha + \gamma)} \cdot ((t^*_o - t^*)^2 - (t_{AB} - t^*)^2).
\]

(3.66)

Differentiating \( TC_A \) with respect to \( t_{BA} \) and \( t_{AB} \), and using Eqs. (3.9), (3.10) and condition (3.31) in the text, one obtains for the cost-minimizing transition times:

\[
t_{BA} = t^* - \frac{\gamma}{2 \cdot (\beta + \gamma)} \cdot \frac{N}{s}, \quad t_{AB} = t^* + \frac{\beta}{2 \cdot (\beta + \gamma)} \cdot \frac{N}{s}.
\]

(3.67)

Substituting Eqs. (3.9), (3.10), and (3.67) into Eq. (3.66) gives

\[
TC_A = \left( \frac{3 + z}{4} - \frac{1}{2} \cdot \sqrt{\frac{\alpha}{\alpha + \gamma} + z^2} \right) \cdot TC^o,
\]

where \( z \equiv \beta \cdot \gamma / ((\alpha + \gamma) \cdot (\beta + \gamma)) \).

**Appendix 3.F** Heterogeneous desired arrival times

3.2.1 Case 1

User A’s fleet costs in Case 1 are:

\[
TC^1_A = \frac{\beta \cdot s}{2} \cdot (t^*_A - t^*_{As})^2 + \frac{\gamma \cdot s}{2} \cdot \left( t^*_A - \frac{N^2_A}{s} - t^*_A \right)^2 = \frac{\beta \cdot \gamma}{2(\beta + \gamma)} \cdot \frac{N^2_A}{s}.
\]
User $B$’s fleet costs are:

$$TC^1_B = \frac{\beta \cdot s}{2} \cdot (t_B^*-t_A)^2 + \frac{\gamma \cdot s}{2} \cdot (t_{Be} - t_B^*)^2$$

$$= \frac{\beta \cdot s}{2} \cdot (t_B^* - t_A^* - \frac{\beta}{\beta + \gamma} \cdot N_A) \cdot (t_B^* - t_B^*)^2 + \frac{\gamma \cdot s}{2} \cdot (t_A^* + \frac{\beta}{\beta + \gamma} \cdot N_A + \frac{N_B}{s} - t_B^*)^2$$

$$= \frac{\beta \cdot s}{2} \cdot \left(\frac{\gamma}{\beta + \gamma} \cdot N_B - x\right)^2 + \frac{\gamma \cdot s}{2} \left(x + \frac{\beta}{\beta + \gamma} \cdot \frac{N_B}{s}\right)^2$$

where $x \equiv t_A^* - t_B^* + (\beta / (\beta + \gamma)) \cdot (N_A/s) + (\gamma / (\beta + \gamma)) \cdot (N_B/s)$. Total system costs are:

$$TC^1 = TC^1_A + TC^1_B = \frac{\beta \cdot \gamma}{2 \cdot (\beta + \gamma)} \cdot s \cdot (N_A^2 + N_B^2) + \frac{\beta + \gamma}{2} \cdot s \cdot x^2.$$ 

The price of anarchy is

$$PA = \frac{TC^1 - TC^o}{TC^o} = \frac{\frac{\beta + \gamma}{4} \cdot s \cdot x^2}{\frac{\beta \cdot \gamma}{2 \cdot (\beta + \gamma)} \cdot (N_A^2 + N_B^2) + \frac{\beta + \gamma}{4} \cdot s \cdot x^2}.$$ 

(3.68)

$PA$ is an increasing function of $x$. Condition (3.49) in the text, $t_B^* - t_A^* \geq (\gamma / (\beta + \gamma)) \cdot (N_B/s)$, implies that $x \leq (\beta / (\beta + \gamma)) \cdot (N_A/s)$. Substituting this value for $x$ into Eq. (3.68), and simplifying, yields

$$PA \leq \frac{\beta \cdot N_A^2}{2 \cdot \frac{\beta}{\gamma} + 2 \cdot \frac{\beta}{\gamma^2} \cdot s}.$$ 

(3.69)

Condition (3.48) in the text, $t_B^* - t_A^* > (\beta / (\beta + \gamma)) \cdot (N_A/s)$, implies that $x \leq (\gamma / (\beta + \gamma)) \cdot (N_B/s)$. This in turn implies $(\beta / (\beta + \gamma)) \cdot (N_A/s) \leq (\gamma / (\beta + \gamma)) \cdot (N_B/s)$, or $N_A \leq (\gamma / \beta) \cdot N_B$. Substituting this inequality into Eq. (3.69) yields

$$PA \leq \frac{\beta}{\gamma} \cdot \frac{\beta}{\gamma^2} \cdot \frac{N_A^2}{N_B^2}.$$ 

(3.70)

In the model variant in Section 3.4.1 the relative magnitudes of parameters $\beta$ and $\gamma$ are unrestricted so there is no upper or lower bound on $\beta/\gamma$. The formula in Eq. (3.70) reaches a maximum at $\beta/\gamma = 1$, for which $PA = 1/5$.

The lower bound on the efficiency index $w$ is derived in the same way as the upper bound on $PA$. Imposing conditions (3.48) and (3.49) as equalities yields $x = (\beta / (\beta + \gamma)) \cdot (N_A/s) = (\gamma / (\beta + \gamma)) \cdot (N_B/s)$. Substituting these equalities into Eq. (3.51) gives

$$w \geq \frac{\beta \cdot \gamma + \beta^2 + \gamma^2}{\frac{3}{2} \beta \cdot \gamma + \beta^2 + \gamma^2}.$$ 

(3.71)

Eq. (3.71) achieves a minimum value of $6/7$ with $\beta = \gamma$. 

77
Chapter 3 Bottleneck model with atomic users

3. F. 2 Case 2

The three transition times in Case 2, \( t_{As}^*, t_{Ac}^*, \) and \( t_{Be}^* \), are solved using three conditions. First, Eq. (3.43) for Case 1 continues to apply, as otherwise user \( A \) could decrease its fleet costs by rescheduling a vehicle from \( t_{As}^* \) to \( t_{Ac}^* \), or vice versa (cf Eq. (3.72a) below). Second, the bottleneck is fully utilized from \( t_{As}^* \) to \( t_{Be}^* \) (cf Eq. (3.72b)). Finally, user \( B \)’s entire fleet must depart during the period \( [t_{As}^*, t_{Be}^*] \) (cf Eq. (3.72c)).

\[
\begin{align*}
\beta \cdot (t_{A}^* - t_{As}^*) &= \gamma \cdot (t_{Ac}^* - t_{A}^*) \quad \text{(3.72a)} \\
\frac{t_{Be}^* - t_{As}^*}{s} &= \frac{N_A + N_B}{(\beta + \gamma) \cdot (\alpha + \beta + \gamma)} \cdot \frac{N_A}{s} \quad \text{(3.72b)} \\
\frac{\alpha}{\alpha + \gamma} \cdot s (t_{Ac}^* - t_{A}^*) + s \cdot (t_{Be}^* - t_{Ac}^*) &= N_B \quad \text{(3.72c)}
\end{align*}
\]

Equations (3.72) resolve to Eqs. (3.52)-(3.54) in the text. Three consistency conditions must be satisfied. First, as in Case 1, user \( B \) cannot gain by rescheduling its last vehicle to \( t_{As}^* \). The requisite condition is:

\[
t_{B}^* - t_{A}^* \geq \frac{\gamma \cdot N_B}{(\beta + \gamma) \cdot (\alpha + \beta + \gamma)} - \frac{\alpha \cdot \beta}{(\beta + \gamma) \cdot (\alpha + \beta + \gamma)} \cdot \frac{N_A}{s} \quad \text{(3.73)}
\]

Second, user \( B \) cannot gain by rescheduling a vehicle at \( t_{A}^* \) to \( t_{Be}^* \). Thus, \( \beta \cdot (t_{B}^* - t_{A}^*) \leq \gamma \cdot (t_{Be}^* - t_{B}^*) \) which reduces to:

\[
t_{B}^* - t_{A}^* \leq \frac{\beta \cdot \gamma}{(\beta + \gamma) \cdot (\alpha + \beta + \gamma)} \cdot \frac{N_A}{s} + \frac{\gamma}{\beta + \gamma} \cdot \frac{N_B}{s} \quad \text{(3.74)}
\]

Condition (3.74) is more stringent than condition (3.34). Third, user \( A \) must stop departing before user \( B \) starts to arrive late (i.e., \( t_{Ac}^* < t_{Be}^* \)) since otherwise the PSNE does not exist unless \( \gamma \leq \alpha \). This follows from the proof of Proposition 3.1. In particular, Lemma 3.1 holds and the only possible PSNE entails queuing. Using a similar reasoning as in Lemmas 3.3 and 3.4 it is possible to show that a PSNE will not exist. Given Eq. (3.53) this implies

\[
t_{B}^* - t_{A}^* \geq \frac{\beta \cdot (\alpha + \gamma)}{\gamma \cdot (\alpha + \beta + \gamma)} \cdot \frac{N_A}{s} \quad \text{(3.75)}
\]

Using Eqs. (3.38) and (3.52), the difference in timing of the social optimum and the PSNE in Case 2 can be written as

\[
t_{As}^* - t_{As}^* = \frac{-\beta \cdot \alpha}{(\beta + \gamma) \cdot (\alpha + \beta + \gamma)} \cdot \frac{N_A}{s} + \frac{x}{2} \quad \text{(3.76)}
\]

Condition (3.74) implies a minimum value of \( x = (\beta \cdot (\alpha + \beta) / ((\beta + \gamma) \cdot (\alpha + \beta + \gamma))) \cdot (N_A/s) \). Condition (3.73) implies a maximum value of \( x = (\beta \cdot (2\alpha + \beta + \gamma) / ((\beta + \gamma) \cdot (\alpha + \beta + \gamma))) \cdot (N_A/s) \). Finally, Condition (3.75) implies a maximum value of \( x = (\gamma / (\beta + \gamma)) \cdot (N_B/s) - (\alpha \cdot \beta^2 / (\gamma \cdot (\beta + \gamma) \cdot (\alpha + \beta + \gamma))) \cdot (N_A/s) \). Applying these values to Eq.
(3.76) yields the feasible range for $t_{2A}^2 - t_{0A}^2$:

$$t_{2A}^2 - t_{0A}^2 \in \left[ \min \left( \frac{-\beta \cdot (\alpha - \beta)}{2 \cdot (\beta + \gamma) \cdot (\alpha + \beta + \gamma)} \cdot \frac{N_A}{s}, \frac{\gamma}{2 \cdot (\beta + \gamma)} \cdot \frac{N_B}{s} - \frac{\alpha \cdot \beta \cdot (\beta + 2\gamma)}{2 \cdot (\beta + \gamma) \cdot (\alpha + \beta + \gamma)} \cdot \frac{N_A}{s} \right) \right].$$

The lower bound applies with $t_B^* - t_A^*$ at its maximum value consistent with condition (3.74). The upper bound applies with $t_B^* - t_A^*$ at its minimum value consistent with conditions (3.73) and (3.75).
Chapter 4

Optimal pricing of flights and passengers at congested airports and the efficiency of atomistic charges
4.1 Introduction

Delays at airports have been consistently increasing over the past years, becoming a major problem worldwide (see, for example, Rupp (2009) and Santos and Robin (2010)). Besides capacity enlargements, the price mechanism has been widely discussed and proposed to manage congestion. This approach involves an airport authority setting user charges, with the possibility to charge passengers, airlines or both. The basic economic motivation for such charges is the congestion externality, as already identified by Pigou (1920) in the context of road traffic. What makes the airport literature different is its focus on congestion pricing when there is market power. This, obviously, introduces a second distortion into the analysis, namely non-competitive pricing. Many papers have studied optimal airport pricing: for example, Brueckner (2002) first showed that in oligopoly, airlines competing in a Cournot fashion internalize congestion imposed on themselves; therefore, the optimal charge should account only for the fraction of congestion that is imposed on competitors. We will refer to this as the “Cournot toll”, as opposed to the “atomistic toll” that considers marginal congestion costs imposed on all flights and passengers, regardless of the operator. One important implication of a Cournot toll is that a dominant airline should pay a lower congestion charge per flight than small airlines (Brueckner, 2005), which is likely to decrease its political feasibility due to distributional issues. Furthermore it would imply that self-financing of airport capacity from the revenues from optimal congestion charges, would become less realistic than in the benchmark case considered originally by Mohring and Harwitz (1962), who showed that with atomistic congestion charges and neutral scale economies in capacity supply, as well as some other technical assumptions, exact self-financing is obtained. Pels and Verhoef (2004) extend the analysis by explicitly considering market power distortions. They show that a welfare maximizing airport has to deal with two inefficiencies: airlines’ market power, that has to be corrected by subsidizing them, and congestion externalities, that require charging the Cournot toll. Further extensions have this congestion pricing rule intact, a consequence of these theoretical works assuming that Cournot competition is representative for airline markets.

However, the Cournot assumption has recently been questioned from the empirical side. Fischer and Kamerschen (2003) estimate airline conduct parameters with U.S. data, finding substantial deviations from Cournot behavior. Fageda (2006) rejects the suitability of the Cournot assumption as representative of the Spanish airline market. Perloff et al. (2007) study airlines’ conduct in a duopoly market using the dataset of Brander and Zhang (1990) and Oum et al. (1993), that has traditionally been used to support Cournot behavior. They, allowing airlines to provide differentiated services on a route, show that
in some routes the outcomes implied by Bertrand behavior are virtually the same as the observed outcomes, while Cournot predictions lie in a less competitive region, not consistent with the data. Finally, Nazarenus (2011) revisit Brander and Zhang’s (1990) study, using data from 2007, concluding that the industry has experienced a regime change from Cournot towards more competitive behavior. This, at least, indicates that other behavioral assumptions, such as Bertrand competition with differentiated products, may be as relevant for aviation markets as the traditional Cournot view.

The purpose of this chapter is to contribute to an ongoing aviation policy debate which focuses on the extent to which airlines indeed internalize congestion effects imposed upon their own flights. Notably, Daniel (1995) and Daniel and Harback (2008) have questioned this feature in Cournot models, while Brueckner and Van Dender (2008) showed how airline behavior approaches atomistic-like non-internalization when there is a Stackelberg leader. Our chapter contributes on a number of respects. First, we derive the optimal pricing policy at congested airports when consumers perceive airlines as imperfect substitutes, and airline behavior follows the Bertrand assumption. Although, as argued above, this case has been shown to be empirically relevant, it has received almost no attention in the theoretical literature. Our aim is to fill this void, by providing the pertinent policy analysis. Second, we distinguish between the role of per-flight and per-passenger tolls for welfare maximization, in order to understand the policy implications of using the one versus the other. To do this, we model the long-run choice of seat capacity as made by airlines, and thus endogenously differentiate between charges per flight and charges per passenger. Third, we assess the relative efficiency of second-best policies, such as atomistic charges, with numerical examples.

We show that, in a duopoly setting where outputs are imperfect substitutes, Bertrand behavior implies that airlines internalize less than the self-imposed congestion, because they take into account the fact that an extra flight imposes congestion on its competitor’s passengers, affecting positively its own demand and profit. This yields an optimal congestion toll that lies between the marginal congestion cost imposed on the competitors’ passengers (the Cournot toll) and the atomistic toll. We also find that the size of the deviation from the self-imposed internalization result of Cournot competition depends on the degree of product substitutability. The entire range of tolls can be optimal: when the substitutability is low, the optimal congestion charge is close to the Cournot toll. Conversely, if the substitutability is high, airlines should be charged a toll very close to the atomistic one, even when they fully recognize the impact of their flight scheduling on airport congestion and even when one of them is a dominant airline with a high market share. We also reproduce the result of Brueckner and Van Dender (2008) for a Stackelberg leader with a Cournot follower, and extend it to the case of a Stackelberg leader with a Bertrand follower, finding optimal tolls that lie between the Cournot toll and the atomistic toll for both players. In this last setting, the optimal toll again approaches the atomistic toll more closely as the substitutability is higher.

Various policy conclusions follow from the analysis. We show that a welfare maximizing airport can only reach the first-best outcome by using two tax instruments, namely per-flight and per-passenger tolls. Moreover, congestion and market power effects are separate:

31 Recall that with differentiated products, the Bertrand outcome no longer entails marginal cost pricing.
32 Empirical evidence for leadership behavior also exists. See for example Morrison and Winston (1995).
the market power exertion can only be corrected by means of a per-passenger subsidy, while the optimal congestion charge should only be charged with a per-flight toll. As a consequence, the welfare maximizing per-passenger toll is below the airport's marginal cost per passenger (due to the subsidy) and the welfare maximizing toll per flight is above airport’s marginal cost per flight (due to the congestion charge). This finding, common to all studied behavioral assumptions, provides important insights on the role of each of the two mentioned instruments and directions on how they should be set. In addition, this result conflicts with a growing tendency of replacing per-movement charges by per-passenger charges, and also with the International Air Transport Association (IATA) position of recovering costs through passenger based charges instead of other aeronautical based charges (IATA, 2010).

We further find that the optimal pricing strategy, in the cases where the Cournot assumption is not representative of the market, includes a congestion charge that is above the marginal congestion cost imposed on the competitors’ passengers, and is likely to be close to the atomistic toll. This has significant policy implications. First, optimal congestion pricing would bring more significant welfare gains and congestion reductions than what has been advanced before on the basis of Cournot assumptions, hence increasing its relevance and efficiency. Second, the degree of self-financing of congested airports would be higher and, in absence of subsidies, it may be close to exact self-financing. Third, the political feasibility of welfare maximizing congestion pricing would be enhanced as the implied differentiation of charges is considerably smaller. For instance, under the Cournot assumption, a firm with 75% market share should pay a congestion toll equal to 25% of the total marginal congestion cost, whereas we find that, considering the behavior and parameters implied by the empirical study of Perloff et al. (2007) for Chicago-based markets, a firm with 75% of market share should pay between 55% and 77% of the total marginal congestion cost. This decreases significantly the potential distributional concerns of optimal congestion pricing.

Finally, we present numerical examples to assess the relative efficiency of second-best policies, and, for example, find that only using a per-flight congestion charge and levying atomistic tolls yield substantial and similar benefits when airlines do not behave in a Cournot fashion, and when the degree of product substitutability is not too low. This complements the findings of Morrison and Winston (2007), who argue in favor of levy atomistic tolls at congested airports, because they find a small net benefit loss when an airport charges the atomistic toll instead of the Cournot toll.

We also believe that our results may help explaining why the empirical and simulation studies provide a wide range of estimations regarding internalization of congestion at airports. In contrast to the road case, where users behave atomistically, the relevant question in aviation markets is what share of congestion airlines actually do internalize when making scheduling decisions of flights. If they internalize a high proportion of congestion costs, charges that optimally account for this will be relatively low and thus should have a small impact on flight patterns and social welfare. The internalization hypothesis, based on Brueckner's analysis, is supported by empirical evidence by Mayer and Sinai (2003) with U.S. data and by Santos and Robin (2010) with European data, who show that delays are lower at highly concentrated airports. On the other hand, Daniel (1995), who first identified the potential for internalization of congestion, argues—with a simulation
model—that atomistic behavior may in fact be more pertinent from an empirical point of view; i.e., that airlines do not take into account self-imposed congestion when making scheduling decisions. As a consequence, the optimal toll should be the so-called atomistic toll that ignores any internalization, and that is equal to total marginal congestion costs. The atomistic behavior of airlines is further supported by empirical evidence by Daniel and Harback (2008) and by Rupp (2009). While the outcome of Cournot competition is in conflict with this evidence, the outcome of the Bertrand setting predicts both a negative relationship between delays and concentration, as well as congestion levels that can be significantly close to the atomistic level.$^{33}$

The chapter is organized as follows. First, in Section 4.2, we introduce the model that includes aircraft size, fare and frequency decisions in an oligopolistic airline market, and that formally takes into account market power exertion and (potential) congestion internalization. In Section 4.3 we derive analytical solutions for the airports’ problem, specifically first-best tolls and optimal capacity investment. Section 4.4 presents numerical exercises to quantify the analytical results, to assess the efficiency of second-best policies, and to study the performance of levying atomistic tolls. Finally, Section 4.5 concludes.

### 4.2 Airlines’ duopoly model

For the analysis, we consider a vertical setting on a single market, i.e., a single origin-destination pair. In the first stage an airport chooses capacity, toll per flight and toll per passenger charged to the carriers that use the facility. In the second stage, a duopoly of airlines compete with aircraft size and frequency as decision variables, in addition to the fare or number of passengers. We choose to analyze analytically a duopoly of carriers in order to keep the simplest and most transparent possible focus on congestion internalization results, the effects of endogenous aircraft size, and the comparison between airlines’ behavioral assumptions; leaving the extension of more than two airlines for the numerical analyses of Section 4.4. Following Zhang and Zhang (2006), we model only one airport for analytical simplicity, but the conclusions remain the same if the other airport is included, as long as the airports share the objective function (this is, in our case, they perform joint welfare maximization).$^{34}$

For the airlines’ market, we consider the differentiated duopoly proposed by Dixit

---

$^{33}$This divergence between views on the extent to which atomistic versus Cournot tolls are more desirable has been addressed before. For example, Brueckner and Van Dender (2008) show that when one airline acts as a Stackelberg leader and interacts with a large number of fringe carriers, the leader behaves atomistically as long as the products are perfect substitutes. Czerny and Zhang (2011) present a different argument and state that, when travelers with different values of time are considered, and in absence of price discrimination by carriers, it might be useful to increase the airport charge towards the atomistic toll. This is to protect passengers with a relatively high value of time from congestion caused by passengers with a relatively low value of time. Obviously, these analyses cannot explain differences in findings between models that consider homogeneous values of time or Nash competition.

$^{34}$If airports are not regulated by the same authority or if airports independently maximize profits both airports have to be formally modeled. For a discussion on the implications of two local welfare maximizing airports see Pels and Verhoef (2004) and for a discussion on independent profit maximization see Basso (2008).
(1979), assuming that demands arise from the following quadratic utility function:

\[ U(q_i, q_j) = A \cdot (q_i + q_j) - (B \cdot q_i^2 + 2 \cdot E \cdot q_i \cdot q_j + B \cdot q_j^2)/2, \]

where \( q_i \) is the amount of good \( i \) (hereafter, when subscript \( j \) appears in the same expression with \( i \), it refers to the rival airline), \( A, B \) and \( E \) are positive parameters with \( B \geq E \geq 0 \) so that goods are imperfect substitutes typically, with the special cases of perfect substitutes occurring when \( B = E \), and independent goods when \( E = 0 \). We consider airlines as imperfect substitutes to account for the fact that not all passengers choose the airline with the most attractive fare-delay combination, therefore allowing airlines with different generalized prices to have passengers in equilibrium. This is motivated by the fact that there are other factors such as loyalty (e.g. due to frequent flyer programs), service levels (e.g. meals and drinks), and consumer preferences for other particular aspects of airlines (e.g. language) that may differ across carriers and make passengers perceive airlines as imperfect substitutes.\(^{35}\)

This utility function gives rise to a linear demand structure, with equivalent inverse and direct demands:

\[ \theta_i = A - B \cdot q_i - E \cdot q_j, \]
\[ q_i = a - b \cdot \theta_i + e \cdot \theta_j, \]

where \( \theta_i \) is the full price of good \( i \) and parameters \( a, b \) and \( e \) satisfy \( a = A/(B + E) \), \( b = B/(B^2 - E^2) \) and \( e = E/(B^2 - E^2) \). Note that the ratio \( e/b \) directly measures the substitutability between airlines, as it ranges from 0 when products are completely independent, to 1 when products are perfect substitutes.

The full price of traveling with airline \( i \) is assumed to be:

\[ \theta_i = p_i + D + g_i. \]

The first term, \( p_i \), is the fare. \( D \) is the passengers’ cost of congestion delays experienced at airports and depends on airport capacity (\( K \)) and on the total number of takeoffs and landings at the congested airport (\( F = f_i + f_j \)). Finally, \( g_i \) is the schedule delay cost faced by a passenger that travels with airline \( i \), which depends only on the flight frequency of the airline (\( f_i \)). The fact that schedule delay does not depend on rival’s frequency, as congestion does, reflects our assumption that in the differentiated duopoly, frequency is perceived as an airline-specific attribute.

We make the plausible assumptions that \( D \) is differentiable in \( F \), that \( g_i \) is differentiable in \( f_i \) and that:

\[ \frac{\partial D}{\partial f_i} > 0, \quad \frac{\partial^2 D}{\partial f_i^2} \geq 0, \quad \frac{\partial^2 D}{\partial f_i \partial f_j} \geq 0, \quad \frac{\partial D}{\partial K} < 0, \quad \frac{\partial^2 D}{\partial K \partial f_i} < 0, \quad \frac{\partial g_i}{\partial f_i} < 0, \quad \frac{\partial^2 g_i}{\partial f_i^2} > 0 \quad \forall i. \]

\(^{35}\)The fact that an airline can have demand despite having a higher generalized price than the competitor has been also modeled with a brand-loyalty variable that gives the additional gain from traveling with a specific airline relative to travel with the other airline (Brueckner and Flores-Fillol, 2007; Flores-Fillol, 2010). In our model it is also possible to model a preference for a specific carrier by letting the demand parameters (\( A, B \) and \( E \)) vary across carriers.
4.2 Airlines’ duopoly model

Congestion thus increases with the number of flights and the marginal effect is stronger when congestion is more severe; congestion decreases with airport’s capacity; schedule delay cost decreases with airline-specific frequency, and that effect is smaller when frequency is higher.\textsuperscript{36}

Following Brueckner (2004), we model airlines’ cost ($C_i$) as a function of aircraft size and frequency in the following way:

$$C_i = f_i \cdot \left( \gamma_i^f + \gamma_i^s \cdot s_i \right),$$

where $\gamma_i^f$ and $\gamma_i^s$ are positive cost parameters and $s_i$ is the number of seats per flight. The underlying assumption is that cost per flight is a linear function of the number of seats, a relation that has been also found in a cost-engineering study for airlines by Swan and Adler (2006).\textsuperscript{37} Congestion costs for airlines are not considered in the analysis because we focus on passengers’ congestion, but including them would not change the results in any essential way.\textsuperscript{38}

With the cost function defined, we can now write the profit of airline $i$ as:

$$\pi_i = q_i \cdot p_i - f_i \cdot \left( \gamma_i^f + \gamma_i^s \cdot s_i \right) - f_i \cdot \tau_i^f - q_i \cdot \tau_i^q,$$

where $\tau_i^f$ is the per-flight toll charged by the airport and $\tau_i^q$ the toll per passenger.

One of the goals of this chapter is to assess the impact of different kinds of strategic interaction on optimal pricing policy at congested airports. In modeling the airlines’ competition, we study the traditional setting for the airlines’ market, namely Cournot competition, where airlines simultaneously choose aircraft size, frequency and number of passengers taking the rivals decision as given. We thereafter look at game with airlines as Bertrand oligopolists, where fare (besides aircraft size and frequency) is the strategic variable instead of quantity. Finally, we study two Stackelberg settings, where the leader chooses all the relevant variables prior to a follower who takes the rival’s number of passengers (output) as given, or the rival’s fare as given.

\textsuperscript{36}This set of assumptions is common in the literature: the linear delay function used by Pels and Verhoef (2004) and the convex function used by Zhang and Zhang (2006) satisfy the assumptions regarding $D$. The schedule delay function that is inversely proportional to the airline frequency satisfies the conditions for $g_i$ (see Brueckner (2004) and Basso (2008) for a discussion).

\textsuperscript{37}In this work, a cost function per trip is calibrated using distance and aircraft size as explanatory variables; then, holding distance fixed, the function is linear in number of seats.

\textsuperscript{38}Congestion imposed on airlines works out in a way similar to congestion imposed on passengers. The intuition is that passenger congestion costs reduce fare that can be charged at given output levels on a dollar by dollar basis. Therefore, the firm weighs “own” passenger congestion costs as heavily as it would weigh “own” congestions costs, and the two types of congestion costs would enter the optimization problem in identical ways. If congestion costs are included, airlines would not internalize congestion costs imposed on the competitors’ flights and this should be corrected in first-best tolls, as found by Basso (2008), Brueckner (2009) and Verhoef (2010).
4.2.1 Cournot behavior

In this game setting, we assume that airlines are Cournot oligopolists in that they choose aircraft size, frequency and number of passengers. Because having idle seat capacity only decreases profit in our model, it is straightforward that an airline will set the number of seats such that the aircrafts are filled (\(s_i = q_i/f_i\)); this allows us to express profit in terms of number of passengers and frequency. Rewriting equation (4.7), using (4.2) and (4.4) we get:

\[
\pi_i = q_i \cdot (A - B \cdot q_i - E \cdot q_j - D - g_i) - f_i \cdot \left(\gamma_i^f + \tau_i^f\right) - q_i \cdot (\gamma_i^s + \tau_i^q),
\]

(4.8)

Then, first-order conditions with respect to number of passengers and frequency yield:

\[
\frac{\partial \pi_i}{\partial q_i} = 0 \Rightarrow p_i = \gamma_i^s + \tau_i^q + q_i \cdot B,
\]

(4.9)

\[
\frac{\partial \pi_i}{\partial f_i} = 0 \Rightarrow \gamma_i^f + \tau_i^f = -q_i \cdot \left(\frac{\partial D}{\partial f_i} + \frac{\partial g_i}{\partial f_i}\right),
\]

(4.10)

Equation (4.9) states that the fare charged by an airline has three terms: (i) the marginal cost per capacity unit (\(\gamma_i^s\)); (ii) the airport charge per passenger (\(\tau_i^q\)); and (iii) a conventional monopolistic markup reflecting carrier’s market power, which is related to the sensitivity of demand and own number of passengers (\(q_i \cdot B\)). Equation (4.10) states that airline’s marginal cost per flight equals marginal benefits for own passengers (marginal congestion savings plus marginal schedule delay benefits); therefore, airlines internalize own-passenger congestion. These rules basically describe that airlines internalize congestion on their own passengers and charge a markup which equals \(q_i \cdot \partial \theta_i/\partial q_i\), a result analogous to the rules obtained previously in Cournot competition (e.g. Pels and Verhoef, 2004; Zhang and Zhang, 2006; Basso, 2008). From now on, to simplify notation, we refer to this game as Cournot competition and we refer to Cournot internalization to the result obtained in this game setting, i.e. perfect internalization of congestion imposed on own-passengers.

It is worth noting at this point that this reduced form of Cournot competition can also be interpreted as a two-stage game where airlines first simultaneously choose aircraft size and frequency and, in the second stage, they compete on fares. As an airline can transport at most \(f_i \cdot s_i\) passengers, in the first stage they are also making a capacity decision, thus the well-known result that a two-stage capacity-constrained competition leads to Cournot outcomes holds.\(^{40}\)

\(^{39}\)The second term in the right-hand side of equation (4.10) is present in previous studies including schedule delay cost. It is in Brueckner (2004) for a monopoly and it is in Basso (2008), but does not appear in the pricing rule of airports because it is set optimally by a private airline from a social welfare perspective.

\(^{40}\)For this result to hold, we only need to assume that when a price-setting firm is capacity constrained, it adjust prices so that the demanded quantity equals its capacity. The intuition comes from the seminal paper of Kreps and Scheinkman (1983), but does not apply directly with imperfect product differentiation. A formal proof and textbook treatment in absence of externalities can be found in Martin (2002), and the formal proof accounting for congestion is available upon request.
4.2 Airlines’ duopoly model

It is interesting to note some aspects about airlines’ behavior that arise from this model. Airlines do not charge passengers directly for congestion because they set frequency and number of passengers separately; for any given demand, they internalize own-passengers congestion by setting frequency according to (4.10) and adjusting aircraft size to accommodate the passengers. In a fixed-proportions model, an additional passenger necessarily increases delays and the only way to internalize this is by charging self-imposed marginal congestion costs to passengers. But, when aircraft size is a strategic variable, this is no longer desirable, because they can accommodate a new passenger, without raising delays, by increasing aircraft size by \( \frac{1}{f_i} \) at a cost of \( \gamma_i^s \) (which they do charge to passengers, see (4.9)). This also explains why the per-flight toll \( (\tau_f^i) \) is absent in the airlines’ fare: it affects frequency and aircraft size setting, while keeping the marginal cost per passenger constant at \( \gamma_i^s + \tau_i^q \).

4.2.2 Bertrand behavior

In this game, the problem faced by an airline is to maximize profit (equation 4.7) with strategic variables being frequency, aircraft size and fare. Again, having idle seat capacity only decreases profit, so that \( s_i = q_i/f_i \) holds and we can rewrite profit in terms of fare and frequency. Rewriting equation (4.7) and explicitly including the functions’ arguments (without including rival’s variables since they are taken as given) we get:

\[
\pi_i(p_i, f_i) = q_i(p_i, f_i) \cdot (p_i - \gamma_i^s - \tau_i^q) - f_i \cdot (\gamma_f^i + \tau_f^i) .
\] (4.11)

Then, first-order conditions with respect to fare and frequency yield:

\[
\frac{\partial \pi_i}{\partial p_i} = 0 \Rightarrow p_i = \gamma_i^s + \tau_i^q + \frac{q_i}{b} , \tag{4.12}
\]

\[
\frac{\partial \pi_i}{\partial f_i} = 0 \Rightarrow \gamma_i^f + \tau_i^q = (p_i - \gamma_i^s - \tau_i^q) \cdot \frac{\partial q_i}{\partial f_i} . \tag{4.13}
\]

Again, the fare charged by an airline includes the marginal cost per capacity unit (\( \gamma_i^s \)), the airport charge per passenger, and a market-power markup. The difference with the Cournot game is that the market power effect is now weaker \( (1/b < b/(b^2 - \epsilon^2) = B) \). The intuition of this comes directly from the type of game: when airlines take rival’s price as given, the outcome is more competitive than when they take rival’s quantity as given (see Singh and Vives (1984) for a discussion in a general context).

From the frequency first-order condition (4.13), we get that frequency is set optimally by equating marginal cost per flight to the revenue gain from an extra flight. For further interpretation, note that using equation (4.3) we can rewrite \( \partial q_i/\partial f_i \) as:

\[
\frac{\partial q_i}{\partial f_i} = -b \cdot \frac{\partial \theta_i}{\partial f_i} + e \cdot \frac{\partial \theta_j}{\partial f_i} = -b \cdot \left( \frac{\partial D}{\partial f_i} + \frac{\partial g_i}{\partial f_i} \right) + e \cdot \frac{\partial D}{\partial f_i} . \tag{4.14}
\]

This expression shows that the effect of an increase in the airline’s number of flights, has two effects on its demand: it changes both schedule delay cost and congestion for own passengers, but it also increases the congestion experienced by competitor’s passengers.
when its frequency is fixed (or taken as given). The second effect has a positive impact for the airline, since increasing the competitor’s congestion raises own demand due to the fact that airlines offer (imperfect) substitute outputs.

Using \( p_i - \gamma_i^s - \tau_i^q = q_i/b \) from equation (4.12), together with equation (4.14), we can rewrite (4.13) as:

\[
\gamma_i f_i + \tau_i f_i = -q_i \cdot \left( \frac{\partial D}{\partial f_i} \cdot \left(1 - \frac{e}{b}\right) + \frac{\partial g_i}{\partial f_i} \right).
\] (4.15)

This equation defines how an airline sets frequency. It differs from equation (4.10) for the Cournot case in the term multiplying marginal congestion costs. In this game, an airline internalizes congestion imposed on its own passengers but also takes into account the congestion imposed on its competitor, as explained above. This is represented by the degree of substitutability \( e/b \) that appears in equation (4.15).

This term causes a difference with the common internalization finding, because now airlines are not taking the competitor’s output as given. In Cournot competition, airlines believe that they are not able to influence the competitor’s number of passengers by raising congestion, simply because they do not “see” the effect by assumption. On the other hand, when taking the competitor’s fare together with frequency as given, output is the result of setting the generalized price through the two variables. Therefore, airlines realize that they can influence the competitor’s output, or increase own demand, by raising the rival’s congestion.

The size of the deviation from the traditional result of internalization depends directly on the degree of substitutability \( e/b \). Recall that this ratio ranges from 0 when products are completely independent to 1 when products are perfect substitutes. The fraction of runway congestion internalized by an airline, when setting frequency, is given by the ratio of congestion terms from equation (4.15) and total marginal congestion costs:

\[
\frac{q_i \cdot D' \cdot (1 - e/b)}{(q_i + q_j) \cdot D'} = \frac{q_i}{q_i + q_j} \cdot \left(1 - \frac{e}{b}\right).
\] (4.16)

As \( e < b \), carriers act as if they internalize less congestion than what is imposed on their own passengers, as in the Cournot model. Only when products are close to be independent, the effective internalization approaches the market share. When they are close to be homogeneous, airline behavior approaches atomistic behavior. For example, if the output is symmetric and the ratio of substitutability is 0.5, airlines internalize only 25 percent of congestion costs, instead of one half.

As in the previous case, the Bertrand reduced form used in this section to represent the airlines market has alternative interpretations. This game setting is equivalent to a two-stage game where airlines first choose aircraft size and, in a second stage, they compete on frequency and fares, as long as they do not directly care about rival’s aircraft size. It is also equivalent to the two-stage game where airlines first choose aircraft size and frequency, and in the second stage they compete on fares, as long as they cannot observe the rival’s actions. In other words, the open-loop equilibrium of the latter two-stage game corresponds to the Bertrand setting analyzed here, and the closed-loop equilibrium to the Cournot setting of Section 4.2.1. From now on, to simplify notation we refer to this game as Bertrand competition and the result regarding internalization in this setting as
4.2 Airlines’ duopoly model

Bertrand internalization.

Which one of the two closed form settings is more appropriate to describe the airline market depends, obviously, on market-specific conditions. Estimations by Brander and Zhang (1990) and Oum et al. (1993)—which have been used to support Cournot behavior—are well summarized by the latter’s conclusion that “the overall results indicate that the duopolists’ conduct may be described as somewhere between Bertrand and Cournot behavior, but much closer to Cournot, in the majority of the sample observations” (p. 189). As their estimations assume perfect substitutability, the Bertrand outcome would correspond to perfectly competitive conditions. Therefore, their conclusion is that fares generally exceed marginal costs, and by less than the “Cournot markup”. This statement is consistent with the Bertrand outcome with differentiated products described above.

In addition, as we discuss in Section 4.1, there are empirical studies that support the internalization hypothesis as well as studies that reject it. We believe that our model helps in explaining such a wide range of findings. As we discuss above, the degree of internalization depends on demand-struc- ture parameters (the ratio of substitutability $e/b$) and, therefore, it is possible that in some markets an airline behaves almost atomistically regarding frequency setting, even if it has a large market share, while in others it internalizes a big share of congestion. Importantly, this still predicts a negative relationship between airport delays and concentration.

Naturally, the suitability of the settings above still remains an empirical question and it can perfectly vary across markets. As suggested by Perloff et al. (2007) with 1980s data, the appropriate setting for the Chicago-Wichita and Chicago-Providence markets is Bertrand competition, analyzed in this Section. In addition, Nazarenus’s (2011) analysis of 37 Chicago-based routes with 2007 data rejects the Cournot hypothesis as representative on average (in contrast to Brander and Zhang (1990)), suggesting that Bertrand behavior with imperfect substitution is more appropriate.\footnote{They reject Cournot behavior as representative in most of the cases, and sometimes the Bertrand behavior as well, because, as they do not consider product differentiation, it is equivalent to reject the perfectly competitive outcome. They conclude that the change of regime is from Cournot towards a more competitive one without reaching perfect competition, which is again consistent with our Bertrand model.}

4.2.3 Stackelberg behavior with a Cournot follower

In the next setting we consider a Stackelberg model, where we suppose that airline $i$ is a leader and airline $j$ the follower that chooses output ($q_j$), frequency ($f_j$) and aircraft size ($s_j$) viewing the leader’s strategic variables ($q_i, f_i$ and $s_i$) as parametric. We refer to this game as Stackelberg-Cournot as a convenient shorthand. As a result of the assumptions, the follower’s behavior is characterized by the first-order conditions in (4.9)-(4.10). The leader maximizes profit knowing the response of the follower to its own decisions. As in previous settings, it is not optimal to have idle capacity, so the profit function of the leader is:

$$\pi_i(q_i, f_i) = q_i \cdot [p_i(q_i, f_i, q_j, f_j) - \gamma_i^s - \tau_i^q] - f_i \cdot (\gamma_if_i^f + \tau_if_i^f),$$

(4.17)
First-order conditions with respect to output and frequency yield:

\[
\frac{\partial \pi_i}{\partial q_i} = 0 \Rightarrow p_i = \gamma_i + \tau_i + q_i \cdot \left( B + E \cdot \frac{\partial q_j}{\partial q_i} + \frac{\partial D}{\partial f_j} \cdot \frac{\partial f_j}{\partial q_i} \right), \quad (4.18)
\]

\[
\frac{\partial \pi_i}{\partial f_i} = 0 \Rightarrow \gamma_i' + \tau_i' = -q_i \cdot \left[ D' \cdot \left( 1 - \frac{\partial f_j}{\partial f_i} \right) + \frac{\partial g_i}{\partial f_i} - \frac{e}{b} \left( \frac{\partial p_j}{\partial f_i} + \frac{\partial g_j}{\partial f_j} \frac{\partial f_j}{\partial f_i} \right) \right] \cdot b, \quad (4.19)
\]

Since the follower’s responses are downward-sloping (the proof is in Appendix 4.B), the leader sets a higher quantity than an airline that takes rival’s output as given (like the follower does). As a consequence, the leader’s market power effect is weaker (for a given frequency, a higher number of passengers implies a lower fare).

As the leader anticipates the way the follower reacts, the incentives to reduce frequency (first-order condition in 4.19) are different from those in the Cournot case, because of two effects. First, the leader predicts that any frequency reduction is partially offset by an increase in the number of flights by the follower \((\partial f_j / \partial f_i < 0)\). As can be seen in (4.19), the term involving marginal congestion is reduced by this expression. This situates the leader’s internalization in between the Cournot case of self-imposed congestion and atomistic behavior, just as pointed out by Brueckner and Van Dender (2008). The second effect—not directly related to marginal congestion—is the last term multiplying \(q_i\) on the right-hand-side of equation (4.19), which further reduces internalization. The leader realizes that any frequency increase induces a reduction on follower’s output, therefore the frequency reduction incentive is diminished. The overall effective internalization is in between the congestion imposed on own passengers and the atomistic behavior. The exact degree, however, depends, among other things, on the degree of substitutability. In Section 4.4, we expand more on this.

### 4.2.4 Stackelberg behavior with a Bertrand follower

Finally, we solve a Stackelberg game with airline \(i\) as a leader and airline \(j\) as a follower that chooses fare \((p_j)\), frequency \((f_j)\) and aircraft size \((s_j)\). The only difference with the previous setting is that the follower takes the leader’s fare as given, instead of output. We refer to this setting as the Stackelberg-Bertrand game.

The problem for the follower is the same as in the Bertrand game, i.e. maximize profit (equation 4.11) with respect to fare and frequency, yielding the same first-order conditions (equations 4.12 and 4.13). For the leader, the profit function now is:

\[
\pi_i(p_i, f_i) = q_i(p_i, f_i, p_j, f_j) \cdot (p_i - \gamma_i - \tau_i) - f_i \cdot (\gamma_i' + \tau_i'). \quad (4.20)
\]

First-order conditions with respect to fare and frequency yield:

\[
\frac{\partial \pi_i}{\partial p_i} = 0 \Rightarrow p_i = \gamma_i + \tau_i + \frac{q_i}{b}, \quad (4.21)
\]

\[
\frac{\partial \pi_i}{\partial f_i} = 0 \Rightarrow \gamma_i' + \tau_i' = -q_i \cdot \left[ D' \left( 1 - \frac{e}{b} \right) + \frac{\partial g_i}{\partial f_i} - \frac{e}{b} \left( \frac{\partial p_j}{\partial f_i} + \frac{\partial g_j}{\partial f_j} \frac{\partial f_j}{\partial f_i} \right) \right] \cdot b, \quad (4.22)
\]
4.3 Airport pricing and capacity investment

where \( \tilde{b} \equiv b - e \cdot \partial p_j / \partial p_i - \partial q_i / \partial f_j \cdot \partial f_j / \partial p_i. \)

In this game, the follower’s responses to an increase of frequency by the leader (\( \partial p_j / \partial f_i, \partial f_j / \partial f_i \)) are downward-sloping while the responses to fare increases (\( \partial p_j / \partial p_i, \partial f_j / \partial p_i \)) are upward-sloping (see Appendix 4.B). This feature prevents us from quantifying analytically whether the market power effect and the congestion internalization are higher or lower in comparison with the Bertrand game of Section 4.2.2. This is because, with general functional forms, the sign of \( \tilde{b} - b \) and the sign of the last term in brackets on the right-hand side of equation 4.22 cannot be determined analytically. In Section 4.4 we study the magnitude of these effects with numerical examples, suggesting that the leader’s internalization can be less than in the Bertrand game, but still in between what is found for Cournot and atomistic behavior.

With the airlines’ market characterized, we now analyze airport pricing and capacity investments.

4.3 Airport pricing and capacity investment

We consider the first-best case of a (unweighted) welfare maximizing airport, with capacity and per-flight as well as per-passenger tolls as instruments. We solve the airport maximization problem analytically in this section, and numerically in Section 4.4. The derivation of first-order conditions presented in this section is in Appendix 4.C.

Social welfare is defined as the sum of net benefits for all agents: consumer surplus, airlines’ profits and airport’s profit. The first of this, with quantities and full-prices being \( q_i, q_j, \theta_i \) and \( \theta_j \), is just \( U(q_i, q_j) - \theta_i \cdot q_i - \theta_j \cdot q_j \). Using (4.1) and (4.2), straightforward calculations yield the following expression:

\[
CS = \frac{B}{2} \cdot (q_1^2 + q_2^2) + E \cdot q_1 \cdot q_2 .
\]

We assume that airport costs are separable and proportional to number of passengers, frequency and capacity, shaping airport profit in the following way:

\[
\Pi = \sum_i q_i \cdot \left( \tau_i^q - c_q \right) + f_i \cdot \left( \gamma_i^f + s_i \cdot \gamma_i^s + c_f \right) - K \cdot r ,
\]

where \( c_q \) is the (constant) operating cost per-passenger, \( c_f \) the (constant) operating cost per-flight, \( r \) the cost of capital and \( K \) the capacity of the airport.

Adding the airlines’ profit (equation 4.7) we can express social welfare as a function of traffic, frequencies, aircraft sizes, and fares:

\[
SW = \left[ \frac{B}{2} (q_1^2 + q_2^2) + Eq_1q_2 \right] + \left[ \sum_i q_i \cdot (p_i - c_q) \right] - \left[ \sum_i f_i \cdot \left( \gamma_i^f + s_i \cdot \gamma_i^s + c_f \right) \right] - [K \cdot r] ,
\]

where the first bracketed term is consumer surplus, the second bracketed term is the airlines’ and the airport’s per-passenger revenues minus costs (the airport’s revenues from tolls cancel out against the airlines’ costs from tolls), the third is per-flight revenues minus
costs (again the airport’s tolls cancel out) and the last bracketed term is capacity costs. Both numbers of passengers are a function of fares and frequencies, but we omit the arguments here.

Straightforward calculations lead to the following conditions for optimal fares:

$$p_i - \gamma_i^s - c_q = 0, \quad \forall i.$$  \hspace{1cm} (4.26)

This result states that optimal fare must equal airline marginal cost per capacity unit plus airport marginal operating cost per passenger. The welfare maximizing fare that should be charged to passengers does not include any congestion term because—as explained in the previous section—airlines take congestion into account only in their frequency setting. From first-order conditions for frequency we obtain the following:

$$-(q_i + q_j) \cdot \frac{\partial D}{\partial f_i} - q_i \cdot \frac{\partial g_i}{\partial f_i} = c_f + \gamma_i^f, \quad \forall i.$$  \hspace{1cm} (4.27)

This means that the optimal frequency must be such that the marginal cost per flight (right-hand side) equals marginal net benefits of all passengers from congestion, plus schedule delay savings (left-hand side). We define “total marginal congestion costs” as the congestion cost that an extra flight imposes on all passengers ($\partial D/\partial f_i \cdot (q_i + q_j)$). If airlines do not internalize any congestion at all, they should be charged this amount plus the airport’s marginal operating cost per flight ($c_f$), the so-called atomistic toll.

Finally, the optimal investment rule for the airport is:

$$-(q_i + q_j) \cdot \frac{\partial D}{\partial K} = r.$$  \hspace{1cm} (4.28)

This shows that airport capacity should be increased until marginal cost equals marginal benefits from congestion reductions. Having established the first-order conditions for social optimal fares and frequencies, we can now derive the optimal tolls per passenger ($\tau_i^q, \tau_j^q$) and per flight ($\tau_i^f, \tau_j^f$), by using the airlines’ first-order conditions for each game (e.g., for Bertrand competition, we use equations (4.12), (4.15), (4.26) and (4.27)). Results now follow, ordered by game type.

- Cournot behavior

For the simultaneous game of Cournot behavior, it can be expected that optimal tolls are consistent with the earlier airport pricing literature. Indeed, the per-passenger toll in (4.29) is marginal operating cost plus a subsidy equal to market power markup, and the per-flight toll in (4.30) equals marginal operating cost plus congestion imposed on the rival airline passengers.

$$\tau_i^q = c_q - q_i \cdot B,$$  \hspace{1cm} (4.29)

$$\tau_i^f = c_f + q_j \cdot D'.$$  \hspace{1cm} (4.30)

However, the subsidy is separate from the congestion toll, and the per-passenger first-best toll is negative when the market power effect is bigger than airport per-passenger marginal operating cost. Hereafter, we use the term “Cournot toll” for the per-flight charge that
4.3 Airport pricing and capacity investment

accounts only for the congestion costs imposed on the competitor’s passengers (equation (4.30)).

- Bertrand behavior

For the simultaneous game of Bertrand behavior, we find:

\[
\tau_i^q = c_q - q_i \cdot \frac{1}{b},
\]  
(4.31)

\[
\tau_i^f = c_f + \left( q_j + \frac{e}{b} \cdot q_i \right) \cdot D'.
\]  
(4.32)

The per-passenger toll in this game is marginal operating cost plus a subsidy equal to the market power markup. For the per-flight toll, we obtain that it is marginal operating cost per flight plus a congestion toll, which is, however, different to the traditional Cournot toll. A welfare maximizing airport charges the congestion costs imposed on the rival’s passengers \((q_j \cdot D')\) plus an additional term. This new term is the own-passenger marginal congestion cost \((q_i \cdot D')\) times the degree of substitutability \((e/b)\). When this ratio is zero, the goods are independent and the optimal per-flight toll is the congestion imposed on the rival’s passengers, as it is in the Cournot game. When it is one, goods are perfect substitutes, and the first-best toll is the so-called atomistic toll. Any other feasible value of \(e/b\) (between zero and one) yields a charge that is somewhere in between the atomistic toll and the Cournot toll. The toll reflects that airlines are internalizing less than self-imposed congestion. Hence, first-best tolls are closer to the atomistic toll than in Cournot competition; and, the higher the degree of substitutability, the higher the first-best toll should be.

The intuition of this result, which to the best of our knowledge is new in the airport pricing literature, is that when goods are independent, there are two monopolies using the same facility in order to serve two independent markets. Therefore, a global welfare maximizing airport charges to each carrier the congestion imposed on the competitor, which was entirely ignored by the operators because of the independence. In the other extreme, where the ratio equals one, goods are perfect substitutes and—since airlines take the rival’s fare as given—the Nash equilibrium is the perfectly competitive outcome (as in Bertrand competition with homogeneous goods). In this extreme case, an airline will expect that any reduction in its own flight volume will be offset by an equally big increase in the competitor’s flight volume. As a consequence, the total number of flights, thus airport congestion, will remain unchanged, and internalization of own congestion makes no sense.

This new result on optimal congestion pricing suggests that welfare maximizing per-flight tolls may be higher than what is suggested by the Cournot model, hence have a more substantial impact in the airlines’ decisions, and therefore may yield more sizable welfare gains. Even if the market is highly concentrated, the dominant airline has to be charged for a high proportion of total marginal congestion costs if the substitutability between the airlines is not too low.
Chapter 4 Optimal pricing of flights and passengers at congested airports

- Stackelberg leader and Cournot follower game

Since the first-order conditions for the follower are the same as in the simultaneous setting with Cournot behavior, the first-best tolling rules also coincide. The per-passenger toll is marginal operating cost plus the market power subsidy (see equation 4.29) and the per-flight toll is marginal operating cost plus congestion imposed on the rival (equation 4.30). For the leader \( i \), the first-best tolls are:

\[
\tau^q_i = c_q - q_i \cdot \tilde{B},
\]

\[
\tau^f_i = c_f + \left( q_j + q_i \cdot \left| \frac{\partial f_j}{\partial f_i} \right| \right) \cdot D' + q_i \cdot E \cdot \left| \frac{\partial q_j}{\partial f_i} \right| \cdot \left| \frac{\partial p_j}{\partial f_i} \right|, (4.34)
\]

where \( \tilde{B} = B + E \cdot \frac{\partial q_j}{\partial q_i} + D' \cdot \frac{\partial f_j}{\partial q_i} \) and derivatives in absolute value are negative (see Appendix 4.C).

The interpretation of these tolls is the same as in previous settings. The first-best per-passenger toll is the marginal operating cost plus a subsidy equal to the market power markup, and the per-flight toll corrects for uninternalized congestion. Because the leader—when considering the effect of its own decisions on the follower’s—alters internalization, the optimal congestion toll charged is in between the Cournot and the atomistic toll. This optimal congestion toll conceptually reproduces the result of Brueckner and Van Dender (2008) in their Stackelberg behavior with a Cournot follower game (note that the last term in (4.34) is not present in their analysis due to the assumptions of perfectly elastic demand and perfect substitution).\(^{42}\)

- Stackelberg leader and Bertrand follower game

The optimal tolling rules for the follower are identical to the tolling rules with Bertrand behavior (equations 4.31 and 4.32). On the other hand, the optimal tolls charged to the leader \( i \) are:

\[
\tau^q_i = c_q - q_i \cdot \frac{1}{b},
\]

\[
\tau^f_i = c_f + \left( q_j + q_i \cdot \frac{b}{b} \right) \cdot D' + q_i \left[ \frac{\tilde{b} - b}{b} \left( D' + g_i \right) + \frac{e - b}{b} D' \cdot \frac{\partial f_j}{\partial f_i} + \frac{e}{b} \left( \frac{\partial p_j}{\partial f_i} + q_j \cdot \frac{\partial f_j}{\partial f_i} \right) \right], (4.36)
\]

where \( \tilde{b} \equiv b - e \cdot \frac{\partial p_j}{\partial p_i} - q_i \cdot \frac{\partial f_j}{\partial f_i} \cdot \frac{\partial f_j}{\partial p_i} \).

The optimal toll per passenger is the airport’s marginal operating cost per passenger plus the subsidy that corrects market power. Since \( \tilde{b} \) is the derivative of own traffic with respect to own fare (taking into account the effect on the follower) the interpretation is the same as usual for pricing with market power. The per-flight toll corrects the frequency setting so that the leader sets the welfare maximizing frequency. As the sign of \( \tilde{b} - b \) cannot be determined a priori, the per-flight toll for the leader has to be studied numerically. In Section 4.4 we do this, finding that in the numerical examples the optimal toll is

\(^{42}\)They relax both assumptions but only in the case of a firm acting as a Stackelberg leader and interacting with a competitive fringe.
in between the Cournot and the atomistic toll, and that it can be above or below the Bertrand toll depending on the degree of substitutability.

For each of the four game types we considered, we found that a welfare maximizing airport needs to use two taxes, namely per-passenger and per-flight tolls, to reach the first-best outcome. It corrects the market power effect with a per-passenger toll and the frequency inefficiency with a per-flight charge. The former \((\tau_q)\) is below airport marginal operating cost per passenger, because it counteracts the airline market power exertion by means of a subsidy. As a consequence, this toll is negative when the airlines’ markups exceed the airport’s marginal operating costs. Conversely, the first-best per-flight toll is always above airport marginal operating cost, because airlines do not fully internalize congestion. If only one tax can be applied, the airport is facing a second-best problem, and which instrument is better to apply depends on market specific conditions. The stronger the airlines’ market power effect compared to the congestion effect, the more likely is that using only per-passenger subsidies is more efficient than charging only per-flight. In the extreme case of monopoly operation, the per-flight congestion toll is unnecessary, and the first-best is attained with per-passenger subsidies only.

The results also imply that the first-best outcome cannot be reached by only charging passengers, because also charges per movement are necessary. If the authority or the facility wants to charge airlines per flight and passengers per trip, a per-passenger tax above the operating costs per passenger is not consistent with welfare maximization. We further expand on this in the numerical analysis below.

Two recent analyses have put a question mark on the desirability of the traditional Cournot congestion toll, i.e. total marginal congestion costs times the market share of each airline (in our model the Cournot per-flight toll in equation 4.30). Morrison and Winston (2007) find a small difference between the net benefits of charging the Cournot toll versus the atomistic toll that ignores any internalization. Then, Brueckner and Van Dender (2008) shows that Stackelberg behavior with a Cournot follower yields optimal airport tolls that lie in between of both policies.

The results of this model give new insights into the debate concerning the desirability of the traditional Cournot congestion toll: first, the optimal toll might well be close to the atomistic toll even without assuming leadership behavior and without abstracting from airlines’ market power exertion. This is the case with Bertrand behavior, simultaneous competition with aircraft size, fare and frequency as strategic variables, and the related two-stage setting. From equation (4.32), it is straightforward to see that the closeness of the optimal toll to the atomistic toll depends on market-specific characteristics (ratio of substitutability \(e/b\)). Therefore, within the same setting but with conditions varying over a network, the entire range of tolls can be optimal. We also confirm the internalization result of Brueckner and Van Dender (2008) for a Stackelberg leader with a Cournot follower, and extend it to the case of a Stackelberg leader with a Bertrand follower, finding even higher optimal tolls. These findings may lead to optimism on the relative efficiency of atomistic congestion pricing in aviation markets. As we cannot compare welfare and equilibrium values analytically, in Section 4.4 we solve numerically the equilibrium for an airport charging the atomistic toll, and make the comparisons with the first-best to assess the relative efficiency.
4.4 Numerical analysis

In this section we present a numerical analysis that allows for making comparisons that are not possible analytically. We also analyze the performance of second-best policies. Despite the simplified structure of the model, we use parameters that are as much as possible calibrated so as to reflect realistic values. We use the following functional forms for the schedule delay cost \( g_i \) and for the passengers’ congestion cost \( D \):

\[
g_i = \gamma \cdot \frac{1}{f_i},
\]

\[
D = \alpha \cdot \left( \frac{f_i + f_j}{K} \right)^\beta,
\]

where the schedule delay cost in (4.37) is inversely proportional to the airline frequency, and \( \gamma \) is a constant representing the monetary value of a unit of schedule delay time. The functional form in (4.37) would be consistent with uniformly distributed desired departure times, and equally spaced flights, and is often used in the literature (e.g. Brueckner, 2004; Basso, 2008). The congestion delay at the airport in (4.38) is a function of the volume capacity ratio, with \( \alpha \) being proportional to the passengers’ value of travel time, \( K \) the capacity, and \( \beta \) the power of the function.

Our reference scenario for calibration has symmetric airlines and assumes a marginal-operating-cost pricing airport, i.e. a toll per passenger of \( c_q \) and a toll per flight of \( c_f \). The parameter calibration considered equilibria in the Cournot and Bertrand settings, and the following tables summarize parameters and equilibrium outputs of the calibration case.

<table>
<thead>
<tr>
<th>Demand parameters</th>
<th>Cost parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airlines</td>
<td>Airport</td>
</tr>
<tr>
<td>A</td>
<td>1250 a 750 ( \gamma ) 30</td>
</tr>
<tr>
<td>B</td>
<td>1.04 b 1.5 ( \alpha ) 26</td>
</tr>
<tr>
<td>E</td>
<td>0.63 e 0.9 ( \beta ) 3</td>
</tr>
</tbody>
</table>

Table 4.1: Parameter values.

<table>
<thead>
<tr>
<th>Game setting</th>
<th>Traffic</th>
<th>Fare</th>
<th>Total Frequency</th>
<th>Aircraft</th>
<th>K</th>
<th>Aggregate Airline Profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bertrand</td>
<td>745</td>
<td>596</td>
<td>2.5</td>
<td>295</td>
<td>3.6</td>
<td>171,452</td>
</tr>
<tr>
<td>Cournot</td>
<td>639</td>
<td>680</td>
<td>2.1</td>
<td>311</td>
<td>3.0</td>
<td>201,493</td>
</tr>
</tbody>
</table>

Table 4.2: Equilibrium outputs, marginal operating cost pricing airport.

We provide numerical examples below for the duopoly setting considered analytically in the previous sections, and also extend the analysis to the case with several airlines. Throughout the analysis, we illustrate possible policy implications of our results by considering the substitutability ratios estimated by Perloff et al. (2007) for the Orlando-Wichita
4.4 Numerical analysis

<table>
<thead>
<tr>
<th>Game setting</th>
<th>Gen. price</th>
<th>Fare share</th>
<th>Congestion share</th>
<th>Schedule delay share</th>
<th>Demand Elasticity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bertrand</td>
<td>629</td>
<td>0.95</td>
<td>0.014</td>
<td>0.036</td>
<td>-1.01</td>
</tr>
<tr>
<td>Cournot</td>
<td>718</td>
<td>0.95</td>
<td>0.011</td>
<td>0.039</td>
<td>-1.35</td>
</tr>
</tbody>
</table>

Table 4.3: Demand elasticities, generalized price and share of generalized price due to fare and delays.

and the Orlando-Providence markets. According to their data, United Airlines and American Airlines had a combined share of passengers above 98% in both markets, and their estimations indicate that the markets are well represented by a Bertrand differentiated duopoly with substitutability ratios between 0.4 and 0.7. Nazarenus (2011), using data from 2007, finds that 37 Orlando based markets fit a duopoly criteria and conclude that the outcome is more competitive than what Cournot behavior implies and less competitive than perfect competition, which is exactly the outcome of differentiated Bertrand competition.43

4.4.1 Internalization

We first solve the problem with a welfare maximizing airport and a duopoly of symmetric airlines, to assess the equilibrium share of congestion that is internalized by each player. In this case, each carrier has an equilibrium share of 50 percent of passengers. Figure 4.1 shows the percentage of congestion that is internalized by each agent for a feasible range of the ratio of substitutability, \( e/b \). Recall that a value of 0 means independent goods, whereas 1 represents pure substitutes. According to the first-order conditions derived in Section 2.2, in the Cournot case each airline internalizes half of the total marginal congestion cost (see equation 4.10). For the Bertrand case, internalization is in between one half and zero, and decreases linearly in the ratio of substitutability; see (4.16). For the game with a Stackelberg leader and a Cournot follower, the leader internalizes less than half of total marginal congestion and the follower acts as in the simultaneous Cournot game, internalizing exactly half of total marginal congestion costs. Finally, Bertrand follower’s internalization is the same as in the simultaneous Bertrand game, while the leader behavior cannot be quantified a priori.

The self-imposed internalization of Cournot behavior and the linear decrease of the internalization for airlines in the Bertrand setting are shown in Figure 4.1 in black lines. To illustrate the possible implications of our results, we may look at parameters estimated by Perloff et al. (2007) for a linear demand structure (as in equation (4.3)) in a duopolistic competition. They find values for the ratio \( e/b \) between 0.4 and 0.7, which implies that airlines would internalize between 12% and 30% of congestion costs, and therefore should

43The duopoly criteria they use is the one proposed by Brander and Zhang (1990), which considers a market as duopolistic when United Airlines and American Airlines together had a market share exceeding 75%, and each carrier had at least 100 passengers in the 10% sample available. Nazarenus (2011) finds that the combined market share of the two airlines is higher that 91% in 29 markets in the third quarter of 2007. Unfortunately the literature on conduct parameters is scarce, and we are not aware of other estimations for the degree of substitutability in the airline industry (following Brander and Zhang (1990), Nazarenus (2011) does not allow for imperfect substitutability).
be charged for a remaining 70% to 88% of the total marginal congestion costs.

For both Stackelberg games, we find that the leader internalization is less than the self-imposed congestion, and decreases towards atomistic behavior as the substitutability as measured by the ratio $e/b$ increases. In the Stackelberg-Cournot case (the two solid lines), the leader always internalizes less than the follower, whose internalization is always the congestion imposed on own passengers. For the Stackelberg-Bertrand setting (the two dashed lines), we find that the leader internalizes roughly the same congestion as the follower, both being less than the self-imposed and approaching to zero as the ratio of substitutability grows. Brueckner and Van Dender (2008) also analyze Stackelberg behavior with a Cournot follower, but suppressing market-power by assuming perfectly elastic demand, and assuming that the outputs of the carriers are perfect substitutes. We extend this to the case of price-sensitive demand and imperfect substitution, finding similar results regarding internalization of congestion: the leader internalizes less than self-imposed congestion, but does not fully reach atomistic behavior. This is represented in Figure 4.1 by the solid gray line.

From this analysis it follows that, for Bertrand behavior both in simultaneous competition as well as in Stackelberg competition, optimal congestion charges are close to the atomistic charge when the substitutability is not too low (near the right-hand end of Figure 4.1). For this reason, optimal per-flight congestion tolls might have a more significant impact on the airlines’ scheduling decisions and, therefore, in alleviating congestion, than what was suggested by earlier Cournot models.

### 4.4.2 Alternative policies

We next study the following alternative policies: (i) the second-best cases where an airport can use only one tax instrument, and (ii) the relative performance of atomistic pricing.
The motivation of studying one instrument is to gain insight on the impact of each tolling instrument separately. The purpose of assessing the performance of atomistic pricing is to better understand to what extent this policy is attractive from an efficiency point of view.

### 4.4.2.1 Using one tax instrument

As shown in Section 4.3, a welfare maximizing airport needs two pricing instruments to reach the first-best outcome. It corrects the market power effect with a per-passenger subsidy and the frequency inefficiency with a per-flight charge. We now look at what happens when it can use only one instrument. For this purpose, we define the relative efficiency \( \Omega^p \) as the welfare gain due to a policy \((p)\), relative to the first-best gain:

\[
\Omega^p = \frac{SW^p - SW^{mc}}{SW^{fb} - SW^{mc}},
\]

where superscript \( fb \) refers to the first-best case, and \( mc \) to an airport charging marginal operating costs (as in the reference scenario for calibration).

Figure 4.2 shows the relative efficiency of both second-best policies for the base parameterization. The results show that, in our calibration, the market power effect dominates the congestion effect, yielding a high relative efficiency for a per-passenger subsidy and a low one for the per-flight toll. The results also show a significant effect of game type on the performance of a policy. As we discuss in Section 2.2, the market power exertion and the amount of internalized congestion are always higher in a Cournot setting than in a Bertrand setting. This leads to a higher relative efficiency of the per-passenger subsidy, and a lower efficiency of the per-flight charge, in the Cournot competition than in the Bertrand competition. For the Stackelberg games, the relative efficiency lies in between the Cournot and Bertrand cases, but the ranking is sensitive to the parameters (as is the degree of internalization). The results furthermore show that, for the chosen parametrization, the social gains for the two instruments are nearly additive. That is, the gains from introducing the one instrument are almost insensitive to the other instrument being in place already. This underlines the lack of substitutability between the two instruments.

The second analysis we perform has the purpose of assessing the relative efficiency of both second-best policies when the market power and the congestion effect have a different comparative importance. We do this by increasing the number of (symmetric) airlines that participate in the market, for a given demand structure, because it captures in one parameter the relative importance of both effects: increasing the number of firms makes the market power effect weaker because of the increased competition and, for the same reason, the congestion externality becomes more severe. As Figure 4.3 shows, the number of firms participating in the market affects the policies in a different manner: the relative efficiency of the per-flight toll increases with the number of airlines, while the opposite occurs for the per-passenger subsidy. The intuition is straightforward: when the number of airlines increases, each airline’s market share of passengers decreases, which leads to less internalization as well as to less ability for exerting market power. Both effects explain the performance improvement of the per-flight toll and the reduction of gains from counteracting the airlines’ markup with a per-passenger subsidy. The negative
relative efficiency of the per-passenger toll, in Bertrand competition for 5 or more airlines, is because congestion inefficiencies are significantly more important than market power exertion in those cases. Therefore, the positive per-flight toll ($\tau_f = c_f$) of the reference scenario, which is removed in the per-passenger toll scenario, gives higher social welfare gains than the second-best per-passenger toll, with a zero per-flight toll.

Which second-best option is better clearly depends on the balance between the inefficiencies and the market structure. The relative efficiency of per-passenger subsidies is higher than the per-flight toll efficiency for Cournot competition with up to 8 airlines; on the other hand, in Bertrand competition, the per-flight toll outperforms the per-passenger subsidy already with 4 airlines, and exceeds 70% of the first-best with 6 airlines.

We illustrated these points by means of increasing the number of firms, but the in-
creasing performance of the per-flight toll in the Bertrand case can also be found with an increase of the ratio of substitutability $e/b$. As the substitutability increases, a smaller share of congestion is internalized spontaneously, and the performance of the per-flight toll rises.\footnote{We assess this numerically, finding that for a ratio $e/b$ of 0.9, $\Omega^f$ reaches 67% for a Bertrand duopoly.}

This exercise provides some useful insights into the performance of second-best policies. When the market power effect is stronger than the congestion effect, it is better for social welfare to give a per-passenger subsidy instead of charging a per-flight toll, and vice versa. If negative tolls are not feasible, it is attractive to charge only a per-flight congestion toll, which will perform better if the internalization is low, because of small market shares, or because substitutability is not too low when airlines behave as in the Bertrand setting. Finally, a positive per-passenger toll cannot be supported from an efficiency perspective, unless the per-flight toll is not feasible and the airlines’ market power markup is small compared to the airport’s marginal operating cost per passenger.

### 4.4.2.2 Atomistic pricing

Finally, we assess the efficiency of levying atomistic congestion tolls to airlines. For this purpose, we look at the welfare gain due to atomistic tolls relative to the second-best case of only having per-flight tolls ($SW^f$), and having a marginal-operating-cost pricing airport as a reference ($SW^{mc}$). The aim of measuring the efficiency relative to this second-best policy, is to isolate the welfare gain that comes from the per-passenger subsidy.\footnote{When the comparison is carried out with respect to the first-best, the welfare gains of charging atomistic tolls are also almost as high as the gains of using the (optimal) per-flight toll. On the other hand, the relative efficiency of atomistic tolls together with (second-best) per-passenger (negative) tolls varies between 99% and 100%. This is because market power effect is dominating.} The performance measure, for this case, is defined in the following way (using superscript $atom$ for atomistic tolls):

$$\tilde{\Omega}^{atom} = \frac{SW^{atom} - SW^{mc}}{SW^f - SW^{mc}}. \quad (4.40)$$

Figure 4.4 shows that $\tilde{\Omega}^{atom}$ has an intuitive relationship with the amount of congestion internalized by carriers. This policy achieves the lowest benefits when airlines behave as Cournot oligopolists, where the first-best toll is half of total marginal congestion costs. Moreover, the performance of atomistic pricing in Cournot competition is only moderately sensitive to the substitutability ratio, as it varies between 77% and 79%.

When the airlines market is characterized by Bertrand behavior, the performance of atomistic pricing rapidly improves as the substitutability is higher ($e/b$ approaches 1). This is because the amount of congestion that is internalized diminishes (see Figure 4.1) and, therefore, the first-best toll approaches the atomistic toll as we showed analytically in Section 4.3. Figure 4.4 also shows that the efficiency measure used ($\tilde{\Omega}^{atom}$) is close to 1 when the degree of substitutability is not too low. Using again the ratios $e/b$ obtained by Perloff et al. (2007), atomistic pricing would yield roughly between 85% and 95% of the maximum social benefit that can be obtained with per-flight tolls.

The benefits that atomistic pricing generates in Stackelberg games follow the inter-
nalization patterns; for the Stackelberg-Bertrand game, $\tilde{\Omega}_{atom}$ is almost the same as in Bertrand competition, because carriers are internalizing approximately the same amount of congestion in both games. For the Stackelberg-Cournot game, we showed, both analytically as well as numerically, that the follower internalizes the same amount of congestion as in the static game with Cournot behavior, and that the leader’s internalization is similar to the one observed in the Bertrand static game. As a consequence, the performance of atomistic pricing, in the Stackelberg-Cournot setting, is lower than in the Bertrand setting, and higher than in the Cournot setting.

In Figure 4.5, we show the relative efficiency of atomistic pricing compared to the first-best (i.e. using $\Omega_{atom}$ from (4.39) as the performance measure), when the number of symmetric firms increases in the Bertrand and Cournot cases. The performance of atomistic pricing, in this case, is also very similar to the performance of the second-best policy of charging only per-flight tolls, as can be seen by comparing Figure 4.3 and Figure 4.5. As the number of airlines increases, the performance of atomistic pricing is better, reaching high values when airlines behavior is well represented by the Bertrand assumption. This implies that atomistic congestion pricing can perform, in terms of social welfare, in a way comparable to the optimal per-flight toll.

These results also give new insights to the airport pricing literature: if congestion is a major issue and the industry is more adequately described as in the Bertrand case, atomistic pricing may offer a more attractive instrument. When per-passenger subsidies are given or, as a second-best policy, per-passenger tolls are set to zero, an airport’s financial deficit can then be reduced without significant welfare losses. If Cournot behavior is more adequate, then naive atomistic pricing for a duopoly can be less attractive. For the Stackelberg games, where first- and second-best tolls differ among carriers, the uniform atomistic toll still produces a small welfare loss with respect to the maximum benefit that
4.5 Conclusions

The present analysis shows that the amount of congestion that airlines internalize may be smaller than the simple market shares formulae from Cournot models, and more so if firms are closer substitutes. As a consequence, the welfare gains and congestion reductions from optimal congestion pricing may be higher. We also show that the airport revenue may be increased and, as a result, a congested airport would be closer to exact self-financing under optimal congestion pricing. Furthermore, the optimal congestion charges may be less differentiated than what has been advanced before, hence diminishing the perception of charges being unequitable and enhancing its implementation feasibility. To what extent these results apply depends on the prevailing type of strategic interaction in a particular market and on the degree of substitutability between airlines. In addition, the chapter differentiates between per-passenger and per-flight charges, showing their unique and non-interchangeable roles in welfare maximization.

We also provide numerical examples to assess the performance of levying atomistic tolls and the relative efficiency of second-best policies. The analysis confirms that when the airlines’ market power effect is larger than the congestion effect, it is wiser to give a per-passenger subsidy instead of a per-flight charge, and vice versa. Numerical examples also suggest that only using a per-flight charge and levying atomistic tolls can yield similar and substantial benefits when the degree of substitutability is not too low, and airlines do not behave in a Cournot fashion. This is, in our framework, when they behave as Bertrand oligopolists, either competing simultaneously (Nash) or in a Stackelberg-Bertrand fashion. The good performance of atomistic pricing, although to a lesser extent, is also found

![Figure 4.5: Relative efficiency for multiple firms relative to first-best.](image)

can be obtained with a per-flight charge.

4.5 Conclusions

The present analysis shows that the amount of congestion that airlines internalize may be smaller than the simple market shares formulae from Cournot models, and more so if firms are closer substitutes. As a consequence, the welfare gains and congestion reductions from optimal congestion pricing may be higher. We also show that the airport revenue may be increased and, as a result, a congested airport would be closer to exact self-financing under optimal congestion pricing. Furthermore, the optimal congestion charges may be less differentiated than what has been advanced before, hence diminishing the perception of charges being unequitable and enhancing its implementation feasibility. To what extent these results apply depends on the prevailing type of strategic interaction in a particular market and on the degree of substitutability between airlines. In addition, the chapter differentiates between per-passenger and per-flight charges, showing their unique and non-interchangeable roles in welfare maximization.

We also provide numerical examples to assess the performance of levying atomistic tolls and the relative efficiency of second-best policies. The analysis confirms that when the airlines’ market power effect is larger than the congestion effect, it is wiser to give a per-passenger subsidy instead of a per-flight charge, and vice versa. Numerical examples also suggest that only using a per-flight charge and levying atomistic tolls can yield similar and substantial benefits when the degree of substitutability is not too low, and airlines do not behave in a Cournot fashion. This is, in our framework, when they behave as Bertrand oligopolists, either competing simultaneously (Nash) or in a Stackelberg-Bertrand fashion. The good performance of atomistic pricing, although to a lesser extent, is also found
numerically in the Stackelberg-Cournot setting.

From the analysis, a number of issues emerge for future research. As airline behavior determines the optimality of congestion charges, and in some settings the degree of substitutability also has a major influence on the size of optimal tolls, estimation of airline conduct parameters stands as an important topic for future research. This is of particular relevance as airlines’ conduct is likely to vary across markets and, in order to implement the correct policy, it is necessary to know which behavior is representative of the market in study. Our approach abstracts from studying entry barriers, which might yield different incentives and outcomes for existent firms, and this framework can be used for studying such potential incentives from a dual tax perspective. Another qualification of our model is that it relies on symmetric product differentiation and, although this does not critically affect our main conclusions, a more realistic demand structure should be considered especially for studying interactions between asymmetric airlines. For example, studying the interactions between legacy and low-cost carriers may require a more elaborate specification of the demand structure.

We see regulation of private airports and the role of commercial (concession) operations in airport pricing as another natural extension of the present analysis. Finally, we perform the analysis in a single market, with one airport for analytical simplicity and to focus on the main insights, but network effects and airports having different objective functions are also seen as a logical extension for future research.
Appendix 4.A  Glossary of notation

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>Inverse demand intercept (reservation price)</td>
</tr>
<tr>
<td>$a$</td>
<td>Demand intercept</td>
</tr>
<tr>
<td>$B$</td>
<td>Inverse demand own-quantity sensitivity parameter</td>
</tr>
<tr>
<td>$b$</td>
<td>Demand own-price sensitivity parameter</td>
</tr>
<tr>
<td>$c_f$</td>
<td>Airport operating cost per-flight</td>
</tr>
<tr>
<td>$c_q$</td>
<td>Airport operating cost per-passenger</td>
</tr>
<tr>
<td>$D$</td>
<td>Passengers’ cost of congestion delays</td>
</tr>
<tr>
<td>$E$</td>
<td>Inverse demand cross-quantity sensitivity parameter</td>
</tr>
<tr>
<td>$e$</td>
<td>Demand cross-price sensitivity parameter</td>
</tr>
<tr>
<td>$e/b$</td>
<td>Degree of substitutability between airlines’ products</td>
</tr>
<tr>
<td>$f_i$</td>
<td>Flight frequency of the airline $i$</td>
</tr>
<tr>
<td>$g_i$</td>
<td>Schedule delay cost faced by a passenger that travels with airline $i$</td>
</tr>
<tr>
<td>$K$</td>
<td>Capacity of the airport</td>
</tr>
<tr>
<td>$p_i$</td>
<td>Fare charged by airline $i$</td>
</tr>
<tr>
<td>$q_i$</td>
<td>Number of passengers traveling with airline $i$</td>
</tr>
<tr>
<td>$r$</td>
<td>Cost of capital</td>
</tr>
<tr>
<td>$s_i$</td>
<td>Aircraft size of firm $i$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Cost of congestion delays parameter</td>
</tr>
<tr>
<td>(proportional to the passengers’ value of travel time)</td>
<td></td>
</tr>
<tr>
<td>$\beta$</td>
<td>Power of the cost of congestion delays function</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>Passengers’ willingness to pay to reduce a unit of schedule delay time</td>
</tr>
<tr>
<td>$\gamma^f_i$</td>
<td>Fixed operating cost per flight of firm $i$</td>
</tr>
<tr>
<td>$\gamma^s_i$</td>
<td>Operating cost per seat per flight of firm $i$</td>
</tr>
<tr>
<td>$\theta_i$</td>
<td>Full price for a passenger traveling with airline $i$</td>
</tr>
<tr>
<td>$\tau^f_i$</td>
<td>Per-flight toll charged to firm $i$</td>
</tr>
<tr>
<td>$\tau^q_i$</td>
<td>Per-passenger toll charged to firm $i$</td>
</tr>
</tbody>
</table>

Table 4.4: Glossary of notation.

Appendix 4.B  Reaction functions

- Stackelberg-Cournot

First-order conditions for the follower $j$ can be written as:

$$A - 2 \cdot B \cdot q_j - E \cdot q_i - D - g_j - \gamma^s_j - \tau^q_j = 0,$$

$$-q_j \cdot \left( \frac{\partial D}{\partial f_j} + \frac{\partial g_j}{\partial f_j} \right) - \gamma^f_j - \tau^f_j = 0.$$
Chapter 4 Optimal pricing of flights and passengers at congested airports

To derive how the follower outputs vary when the leader changes quantity and frequency, we differentiate the system, write the result in matrix notation, and apply Cramer’s rule. After some straightforward calculations, we get:

\[ \frac{\partial q_j}{\partial q_i} = -\frac{E \cdot q_j \cdot (D'' + g''_j)}{R} \leq 0, \]  
\[ \frac{\partial f_j}{\partial q_i} = \frac{E \cdot (D' + g'_j)}{R} \leq 0, \]  
\[ \frac{\partial q_j}{\partial f_i} = -\frac{q_j \cdot (D' \cdot g''_j - D'' \cdot g'_j)}{R} \leq 0, \]  
\[ \frac{\partial f_j}{\partial f_i} = -\frac{2 \cdot B \cdot q_j \cdot (D'' + D' + g'_j)}{R} \leq 0, \quad (4.43)\]

where \( R = 2 \cdot B \cdot q_j \cdot (D'' + g''_j) - (D' + g'_j)^2 \) is, by definition, the determinant of the Hessian matrix of airline profit, and, since we assume the existence of a maximum, \( R > 0 \). Because of assumptions in (5), and the fact that an equilibrium with positive traffic implies \( D' + g'_j < 0 \), it is clear that all the reaction functions in this case are non-positive.

- Stackelberg-Bertrand

We proceed in the same way as above to calculate the reaction functions. First-order conditions for the follower \( j \) are:

\[ p_j - \gamma^q_j - \tau^q_j - \frac{q_j}{b} = 0, \]  
\[ \gamma^f_j + \tau^f_j - (p_j - \gamma^q_j - \tau^q_j) \cdot \frac{\partial q_j}{\partial f_j} = 0. \]  

And denoting \( \hat{p}_j = p_j - \gamma^q_j - \tau^q_j \), we obtain:

\[ \frac{\partial p_j}{\partial p_i} = -\frac{e}{b} \cdot \frac{\partial^2 q_j}{\partial f_j^2} \cdot \hat{p}_j 
\geq 0, \]  
\[ \frac{\partial p_j}{\partial f_i} = -\frac{1}{b} \cdot \frac{\partial q_j}{\partial f_j} \cdot \frac{\partial^2 q_j}{\partial f_j^2} \cdot \hat{p}_j + \frac{1}{b} \cdot \frac{\partial q_j}{\partial f_j} \cdot \frac{\partial^2 q_j}{\partial f_j^2} \cdot \frac{\partial q_j}{\partial f_j} \cdot \frac{\partial q_j}{\partial f_j} \cdot \hat{p}_j \leq 0, \]  
\[ \frac{\partial f_j}{\partial p_i} = \frac{e}{b} \cdot \frac{\partial q_j}{\partial f_j} \geq 0, \]  
\[ \frac{\partial f_j}{\partial f_i} = \frac{2 \cdot \partial^2 q_j}{\partial f_j \partial f_j} \cdot \hat{p}_j + \frac{1}{b} \cdot \frac{\partial q_j}{\partial f_j} \cdot \frac{\partial q_j}{\partial f_j} \leq 0, \]  

where \( R = -2 \cdot \frac{\partial^2 q_j}{\partial f_j^2} \cdot \hat{p}_j - \frac{1}{b} \cdot (\frac{\partial q_j}{\partial f_j})^2 > 0, \) as we assume existence of a maximum. Because
of assumptions (5), \( \partial q_j / \partial f_i < 0, \partial^2 q_j / \partial f_i^2 < 0 \) and \( \partial^2 q_j / \partial f_i \partial f_j < 0 \). An equilibrium with positive traffic implies \( \partial q_j / \partial f_j > 0 \), therefore we can show that reaction functions have the sign that is presented above.

**Appendix 4.C Welfare maximizing airport first-order conditions**

Since airlines choose aircraft size such that \( q_i = s_i \cdot f_i \), we use this relation before maximizing welfare. Then, first-order condition with respect to fare is:

\[
\frac{\partial SW}{\partial p_i} = B q_i \frac{\partial q_i}{\partial p_i} + B q_j \frac{\partial q_j}{\partial p_i} + E q_i \frac{\partial q_i}{\partial p_i} + E q_j \frac{\partial q_j}{\partial p_i} + (p_i - \gamma_i^s - c_q) \cdot \frac{\partial q_i}{\partial p_i} + q_i (p_j - \gamma_j^s - c_q) \cdot \frac{\partial q_j}{\partial p_i}
\]

\[
= q_i \left( B \frac{\partial q_i}{\partial p_i} + E \frac{\partial q_j}{\partial p_i} + 1 \right) + q_j \left( B \frac{\partial q_j}{\partial p_i} + E \frac{\partial q_i}{\partial p_i} \right) + (p_i - \gamma_i^s - c_q) \cdot \frac{\partial q_i}{\partial p_i} + (p_j - \gamma_j^s - c_q) \cdot \frac{\partial q_j}{\partial p_i} = 0.
\]

And noting that

\[
B \frac{\partial q_i}{\partial p_i} + E \frac{\partial q_j}{\partial p_i} = B \cdot -b + E \cdot e \frac{b^2}{b^2 - c^2} + \frac{c^2}{b^2 - c^2} = -1,
\]

\[
B \frac{\partial q_j}{\partial p_i} + E \frac{\partial q_i}{\partial p_i} = B \cdot e + E \cdot -b \frac{b \cdot e}{b^2 - c^2} + \frac{-b \cdot e}{b^2 - c^2} = 0,
\]

we obtain, using the analogous calculations for \( \partial SW/\partial p_j \), first-order conditions for fares:

\[
\frac{\partial SW}{\partial p_i} = (p_i - \gamma_i^s - c_q) \cdot \frac{\partial q_i}{\partial p_i} + (p_j - \gamma_j^s - c_q) \cdot \frac{\partial q_j}{\partial p_i} = 0,
\]

\[
\frac{\partial SW}{\partial p_j} = (p_i - \gamma_i^s - c_q) \cdot \frac{\partial q_i}{\partial p_j} + (p_j - \gamma_j^s - c_q) \cdot \frac{\partial q_j}{\partial p_j} = 0.
\]

Both conditions can only be satisfied if fares fulfill:

\[
(p_i - \gamma_i^s - c_q) = 0, \forall i.
\]

For the frequency first-order conditions, let \( \overline{p_i} = p_i - \gamma_i^s - c_q \) and \( \overline{p_j} = p_j - \gamma_j^s - c_q \). Then, the derivative of SW with respect to \( f_i \) is:

\[
\frac{\partial SW}{\partial f_i} = B q_i \frac{\partial q_i}{\partial f_i} + B q_j \frac{\partial q_j}{\partial f_i} + E q_i \frac{\partial q_i}{\partial f_i} + E q_j \frac{\partial q_j}{\partial f_i} + \overline{p_i} \cdot \frac{\partial q_i}{\partial f_i} + \overline{p_j} \cdot \frac{\partial q_j}{\partial f_i} - (\gamma_i^f + c_f).
\]

Using \( \overline{p_i} = \overline{p_j} = 0 \) from equation (4.26), we can write,

\[
\frac{\partial SW}{\partial f_i} = q_i \left( B \frac{\partial q_i}{\partial f_i} + E \frac{\partial q_j}{\partial f_i} \right) + q_j \left( B \frac{\partial q_j}{\partial f_i} + E \frac{\partial q_i}{\partial f_i} \right) - (\gamma_i^f + c_f).
\]
Chapter 4 Optimal pricing of flights and passengers at congested airports

Note that,

\[
\frac{\partial q_i}{\partial f_i} = -b \frac{\partial D}{\partial f_i} - b \frac{\partial g_i}{\partial f_i} + e \frac{\partial D}{\partial f_i} ,
\]
(4.61)

\[
\frac{\partial q_j}{\partial f_i} = -b \frac{\partial D}{\partial f_i} + e \frac{\partial D}{\partial f_i} + e \frac{\partial g_i}{\partial f_i} .
\]
(4.62)

And using these equations, first-order condition can be written as,

\[
\frac{\partial SW}{\partial f_i} = - (\gamma_i^f + c_f) + q_i \left( \frac{\partial D}{\partial f_i} (-bB + eE) + \frac{\partial D}{\partial f_i} (eB - bE) + \frac{\partial g_i}{\partial f_i} (-bB + eE) \right) \\
+ q_j \left( \frac{\partial D}{\partial f_i} (-bB + eE) + \frac{\partial D}{\partial f_i} (eB - bE) + \frac{\partial g_i}{\partial f_i} (eB - bE) \right).
\]
(4.63)

Finally, because \(-b \cdot B + e \cdot E = -1\) and \(e \cdot B - b \cdot E = 0\), the first-order condition is:

\[
-(q_i + q_j) \cdot \frac{\partial D}{\partial f_i} - q_i \cdot \frac{\partial g_i}{\partial f_i} = c_f + \gamma_i^f .
\]
(4.64)

The derivative of social welfare with respect to capacity is:

\[
\frac{\partial SW}{\partial K} = Bq_i \frac{\partial q_i}{\partial K} + Bq_j \frac{\partial q_j}{\partial K} + Eq_i \frac{\partial q_i}{\partial K} + Eq_j \frac{\partial q_j}{\partial K} + \overline{p_i} \cdot \frac{\partial q_i}{\partial K} + \overline{p_j} \cdot \frac{\partial q_j}{\partial K} - r .
\]
(4.65)

Noting that,

\[
\frac{\partial q_i}{\partial K} = -b \frac{\partial D}{\partial K} + e \frac{\partial D}{\partial K} , \quad \forall i ,
\]
(4.66)

and using \(\overline{p_i} = \overline{p_j} = 0\) from equation (4.26), \(-b \cdot B + e \cdot E = -1\) and \(e \cdot B - b \cdot E = 0\), we can write first-order condition for capacity as:

\[
-(q_i + q_j) \cdot \frac{\partial D}{\partial K} = r .
\]
(4.67)
Chapter 5

Input third-degree price discrimination by congestible facilities
5.1 Introduction

Congestible facilities often provide infrastructure access to downstream firms that may operate in different markets. For example, access to the airport’s runway is an essential input for an airline’s production in multiple city-pairs. In many countries, input price discrimination is banned by law, so that suppliers must charge uniform prices to firms. To a large extent the ban on input price discrimination applies also to congestible facilities. For example, the EU Airport Charges directive (2009/12/EC) prohibits differentiated charges to airlines using the same service (i.e. terminal and level of service). A similar ban holds for airports in the U.K. (Section 41 of the 1986 Airports Act) and in the U.S. (2013 FAA’s Policy Regarding Airport Rates and Charges). The regulations of the World Trade Organization (WTO) through the General Agreement on Tariffs and Trade (GATT) basically do not allow price discrimination by ports. Similar examples can be found in other transport sectors too. It is therefore evident that current bans on price discrimination to congestible facilities have an impact on many large economic sectors.

Congestible facilities feature two characteristics that make the analysis distinct from the traditional price discrimination studies in input markets. First, there is congestion: an output increase by one firm imposes additional costs on consumers of all markets, therefore reducing the price that other firms can charge. Congestion, thus, makes demands interrelated in a way analogous to substitution. In addition, downstream firms do not fully internalize this externality, so that the aggregate output may be inefficiently high. Second, the ownership form of the congestible facilities subject to price discrimination regulation is diverse: for example, in Europe alone, the ban applies to private, public and mixed private-public airports. The incentives of the facilities to apply price discrimination and, therefore, its effect on welfare may vary with the ownership form. The purpose of this chapter is to study third-degree price discrimination by both private (profit maximizing) and public (domestic welfare maximizing) congestible facilities, and shed light on whether and when a broad (e.g. Europe-wide) ban on input price discrimination is desirable.

In order to focus on the effects of congestion and ownership form, we analyze a case with two downstream markets and two downstream firms, one domestic and one foreign, that are equally efficient. Each firm is a monopoly in one market and the only interdependency is through congestion. To highlight the differences with the uncongested case, we consider an industry structure that is comparable to the commonly used structure in the literature in that the input provider is a monopolist and firms take the input price as given. A private facility would therefore differentiate charges according to the different demand conditions in the two markets. A public facility also considers domestic consumer surplus

\footnote{A version of this chapter will appear in the Tinbergen Institute Discussion Paper series as Silva (2015). I am grateful to Vincent van den Berg, Achim Czerny and to Erik Verhoef for their helpful comments. The main variables and parameters that are used in this chapter are summarized in Appendix 5.A.}

\footnote{The ban applies to the airport with the highest passenger movement in each EU Member State and to any airport whose annual traffic is over 5 million passengers.}

\footnote{The FAA’s Policy Regarding Airport Rates and Charges prohibits “unjust discrimination”. This prohibition does not prevent airports to set different charges to different aeronautical users (such as signatory and nonsignatory carriers) or to in peak and off-peak periods. Nevertheless, it explicitly bans, for example, differentiated charges to firms that belong to the same category irrespective of the markets they serve, and to foreign and domestic airlines engaged in similar international air services.}
and the profit of the domestic firm, so it would give price concessions to the domestic firm in detriment of the foreign firm, to stimulate (domestic) production and capture foreign profit.

In this chapter we, first, analyze the price, output and welfare effect of input third-degree price discrimination by a private facility and assess how the presence of congestion externalities affects the analysis. Second, we study the case of a public facility and assess the impact of the ownership form on the effects of price discrimination. Finally, we compare the welfare effect of price discrimination under both ownership forms and elaborate on the desirability of a broad ban on price discrimination.

Input third-degree price discrimination when downstream firms are equally efficient and operate in multiple markets has been recently explored by Arya and Mittendorf (2010). In a two-market setting and using linear demands, they show that the aggregate output is the same under both pricing regimes (i.e. with and without discrimination), but price discrimination leads to an output shift from the market with higher demand to the market with lower demand. Therefore, extending the intuition from final good markets, price discrimination leads to welfare deterioration in the case where there is only one firm operating in each market.\textsuperscript{49,50} Our chapter contributes to this branch of the literature by studying the case with interrelated demands through congestion and by considering a domestic-welfare-maximizing input provider. Also with linear demands, yet in the presence of congestion, we find benefits from price discrimination when the willingness to pay to reduce travel delays differs across markets. This may be due to different average income of consumers or different composition of trip purpose (leisure versus business) across markets, among others. We show that under private and public ownership, input price discrimination can increase aggregate output and welfare. This result suggests that the presence of congestion externalities enlarges the extent to which input price discrimination is desirable.

The literature on price discrimination under negative consumption externalities has mainly focused on final markets. Adachi (2005), considering only consumption externalities within markets, shows that welfare can increase when third-degree price discrimination is allowed when output does not. Czerny and Zhang (2015) study price third-degree discrimination by a monopoly airline considering cross- and within-market negative externalities together, a feature that is typical of congestion. They show that there is a time-valuation effect of price discrimination that works in the opposite direction as the output effect and, as a result, welfare can increase when output decreases. When demands are linear, they find that price discrimination reduces the aggregate passenger quantity, which reduces congestion costs, and that this can increase welfare. We show that price discrimination in input markets under congestion externalities exhibits fundamental differences with the case of final markets and that the analysis provides essentially different

\textsuperscript{49} They also analyze the case in which a firm operates in both markets and faces different degrees of competition in each one. When the market with lower demand is also the market with lower competition, the increased production incentives under price discrimination in this market may increase welfare.

\textsuperscript{50} There is also a large stream of literature studying third-degree price discrimination in input markets where downstream firms have different levels of (cost) efficiency that shows benefits of uniform pricing (e.g., Katz, 1987; DeGraba, 1990; Yoshida, 2000; Valletti, 2003). Nevertheless, uniform pricing can be harmful when there is bargaining between buyers and suppliers (O’Brien and Shaffer, 1994), and when there is input demand-side substitution (Inderst and Valletti, 2009).
Chapter 5 Input third-degree price discrimination by congestible facilities

insights. We find that under linear demands, input third-degree price discrimination by a profit maximizing facility can yield higher total welfare and consumer surplus than what is obtained under uniform pricing by leading to aggregate output expansion and to a reduction of both prices. As congestion effects work in a similar way as the substitution effect, the intuition is similar to the one provided by Layson (1998) for substitute final goods. Reducing the price in one market can reduce the profitability of the other and, if this effect is large enough, it can cause that the price in the other market has to also be reduced. These results are in sharp contrast with the outcome of the models in final markets. The difference arises because the input provider faces derived demands, which may have essential differences with final good demands. For example, under price discrimination by a private supplier, the market with the highest input price can be the market with the lowest final good price. Therefore, what could be called the “weak” market in terms of final price can be the “strong” market for the input provider.

Our analysis also contributes to the transport policy literature. Benoot et al. (2013) study price discrimination by a local welfare maximizing airport when passengers are homogenous and airlines and markets are symmetric. The incentives for price discrimination arise from that the foreign passengers’ surplus is not fully considered. They numerically find that welfare is higher under uniform pricing because foreign passengers surplus increases.51 We focus on the more fundamental question of whether and when a broad ban on price discrimination is welfare enhancing. We find that under uniform pricing total welfare may be higher than under price discrimination by a domestic welfare maximizing facility also when there is asymmetry, but this is not the only possible outcome as price discrimination can increase welfare. Importantly, we find that under the conditions that make uniform pricing by a public facility welfare enhancing, price discrimination may yield higher total welfare than uniform pricing if the input supplier is private. We also find that the reverse may happen: price discrimination by a public facility can increase total welfare under the same conditions that make uniform pricing socially optimal if the supplier is private. These results have important policy implications. The ownership form of transport facilities has been consistently moving from public to private in the last decades; for example, in 2010, 48% of all European traffic was handled by a fully privatized airport or by mixed private-public airports. Our results suggests that a ban on price discrimination that covers a large number of congestible facilities and, in particular, that covers different ownership forms has to be revised, especially in the light of the privatization wave.

The remainder of the chapter is structured as follows. Section 5.2 introduces the model and main assumptions. Section 5.3 analyzes the effects of price discrimination by a private facility while Section 5.4 analyzes the case of a public facility. Section 5.5 compares the welfare effect of price discrimination under both ownership forms. Section 5.6 extends the conclusions to downstream perfect price discrimination and Section 5.7 concludes.

51Haskel et al. (2013) study price discrimination by substitute private airports when airlines and markets are symmetric. In their model, the incentives to price discriminate arise from that there is bargaining between airports and airlines, so that the possibility of differentiated prices changes the bargaining structure. Their main result is that price discrimination leads to lower prices as it makes airlines “tougher” negotiators.
5.2 The model and the downstream markets

We study price discrimination by a monopolist transport facility that sells access to its infrastructure, which is an input necessary for downstream production. There are two downstream markets served by the facility, $A$ and $B$, which may represent movement of people or cargo to different destinations. Markets are interrelated through congestion as an additional unit of output in any market imposes an externality on all other consumers, but are otherwise independent. Downstream firms transform one unit of input into one unit of output. Think, for example, of an airport setting per-passenger charges to airlines flying to different cities, where congestion occurs at the passenger’s facilities (security passenger and baggage screening or access to gateways) and/or on the runway as a result of aircraft landing and take-off.

There are two downstream firms and we denote them in the same way as the markets in which they operate. Thus, firm $i$ operates in market $i$ with $i = \{A, B\}$. Each firm’s demand $q_i$ depends on the full price faced by consumers, which is the sum of the downstream firm’s price (e.g. ticket) and the cost of congestion (e.g. delays at the airport). The delay due to congestion, $D(Q)$, increases in the aggregate consumption ($Q = q_A + q_B$), to reflect within- and cross-market negative consumption externalities. As every unit of output causes delays on all others, a natural interpretation for the market demand is that it is the aggregation of consumers who buy either 0 or 1 unit of the good (e.g. a trip in a peak period) and are heterogeneous in their willingness to pay for the good. Denote $P_i(q_i)$ the downward-sloping inverse demand in market $i$ and $v_i$ the willingness to pay to reduce congestion delays, or the value of time as shorthand, which is assumed to be the same for all individuals in a market but to be different across markets. Without loss of generality we assume that consumers in market $A$ have a higher time valuation than consumers in market $B$, so that $v_A > v_B$ holds in the remainder of the chapter. We also assume that downstream firms have constant marginal costs and, following Singh and Vives (1984), that their costs are incorporated through the intercept of the inverse demand function.

In the analysis that follows, we study the case where downstream firms set a unit price and cannot discriminate consumers so, in equilibrium, the firm’s price equals $P_i(q_i) - v_i \cdot D(Q)$, the marginal willingness to pay net of congestion delay costs. Section 5.6 extends the analysis by relaxing this assumption and studies downstream firms applying first-degree price discrimination. Consequently, for a given input price, $w_i$, the downstream firm $i$ maximizes:

$$\pi_i = q_i \cdot \left[ P_i(q_i) - v_i \cdot D(Q) - w_i \right],$$  \hspace{1cm} (5.1)

and the first-order condition leads to the following pricing rule:

$$\frac{\partial \pi_i}{\partial q_i} = 0 \Rightarrow P_i(q_i) - v_i \cdot D(Q) = w_i + q_i \cdot \left[ -P_i'(q_i) + v_i \cdot D'(Q) \right].$$  \hspace{1cm} (5.2)

52We assume that congestion does not affect the downstream firms’ costs, but this could be readily included in our analysis without changing the main results and conclusions. The reason is that congestion does affect firms in that increased congestion raises the full price faced by consumers and therefore final good prices will be lowered by the increased congestion. In the downstream firms’ profit function, whether congestion raises the costs or reduces the passengers’ willingness to pay makes no difference.

53If $a_i$ is the inverse demand intercept in market $i$ and $c_i$ the marginal cost, we may replace $A_i$ by $a_i - c_i$. 

115
Chapter 5 Input third-degree price discrimination by congestible facilities

Eq. (5.2) shows that the firm’s pricing rule has the facility charge \( (w_i) \), a traditional monopoly market power markup \((-q_i \cdot P_i')\) and the marginal congestion cost on firm \( i \)’s own consumers \((q_i \cdot v_i \cdot D')\). The downstream firm realizes that an additional consumer raises congestion and reduces the price it can charge, but does not internalize the effect on the other firm’s consumers. This internalization result was first recognized by Daniel (1995) in the context of airport congestion pricing and explored theoretically by Brueckner (2002). The system of first-order conditions in Eq. (5.2) for both firms defines the derived demands faced by the input provider \( q_A(w_A, w_B) \) and \( q_B(w_A, w_B) \). The closed form for the derived demands are in Appendix 5.B.

Before moving into the supplier’s maximization problem, it is useful to compare the downstream price with the welfare maximizing price. In this model, total welfare is:

\[
W = \sum_i \left[ \int_0^{q_i} P_i(x)dx \right] - \left[ \sum_i v_i \cdot q_i \right] \cdot D(Q),
\]

(5.3)

and the welfare maximizing downstream pricing rule is

\[
\frac{\partial W}{\partial q_i} = 0 \Rightarrow P_i(q_i) - v_i \cdot D(Q) = [q_A \cdot v_A + q_B \cdot v_B] \cdot D'(Q) \quad \forall \; i \in \{A, B\}.
\]

(5.4)

A comparison between Eq. (5.2) and Eq. (5.4) reveals that, input prices aside, the prices set by downstream firms are not necessarily higher than optimal. If the demand is sufficiently elastic, i.e. the demand-related markup is low compared to the un-internalized externality (e.g. \(-q_A \cdot P_A' < q_B \cdot v_B \cdot D'\)), prices will be too low and output too high. This result and its implications for airport pricing have also been discussed in the air transportation literature (see e.g. Pels and Verhoef, 2004).

As one of our aims is to analyze the role of congestion on the effects of input price discrimination, we follow much of the literature and assume that demands are linear. This allows us to compare our results to those in the previous literature more transparently, as mainly models with linear demands have been used to study the effect of price discrimination in final goods markets under negative consumption externalities (e.g. Adachi, 2005) and when input buyers participate in multiple markets (e.g. Arya and Mittendorf, 2010). We also assume, for simplicity, that the congestion delay function is linear in the aggregate quantity. Note that assuming linear functional forms does not mean that we confine our analysis to a constant aggregate output, because in presence of within- and cross-market congestion externalities and linear demands, the output effect of price discrimination is not zero when time valuations are different (Czerny and Zhang, 2015).

The pricing regimes that we study are uniform pricing, where the facility is restricted to charge all firms the same price per unit of output, and price discrimination, where the facility is allowed to charge different unit prices.\(^{54}\) We assume throughout the chapter that all markets are always served under both pricing regimes. The equilibrium concept that

\(^{54}\)There is a distinction between price differentiation and price discrimination in congestible markets (see e.g. van der Weijde, 2014). As in our setting the marginal external cost \((\sum_i v_i \cdot q_i \cdot D'(Q))\) is the same for all consumers, there is no difference between discrimination and differentiation, so we use them interchangeably.
we use is subgame-perfect Nash equilibrium, and we use backward induction to identify it. We first study the case of a profit maximizing facility.

5.3 Private facility

5.3.1 Price discrimination

When price discrimination is allowed, the facility chooses $w_A$ and $w_B$ to maximize:

$$
\Pi^{PD} = w_A \cdot q_A(w_A, w_B) + w_B \cdot q_B(w_A, w_B) .
$$

(5.5)

where we normalize the input supplier’s costs to zero. The first-order conditions lead to the closed-form solutions for $w_A$ and $w_B$ (see Appendix 5.B) and imply the following pricing rules:

$$
w_A = 2 \cdot q_A \cdot \left[ -P'_A + v_A \cdot D' \right] + q_B \cdot v_B \cdot D',
$$

(5.6)

$$
w_B = 2 \cdot q_B \cdot \left[ -P'_B + v_B \cdot D' \right] + q_A \cdot v_A \cdot D'.
$$

(5.7)

Not surprisingly, the input provider also exerts market power and consumers face a double marginalization. In addition, the facility charges the marginal congestion cost that is not internalized by the firm (the last term on the right-hand side of Eqs. (5.6) and (5.7)). Therefore, under price discrimination, the final price in each market is higher than the socially optimal price and output is inefficiently low. This result is useful for the welfare analysis below and it is essentially different to the case of final good markets and congestion externalities where the quantity under price discrimination can be inefficiently high. This is because the downstream firm’s markup is not necessarily higher than the marginal external congestion cost, but the sum of the downstream and upstream markups is.

In our analysis a crucial aspect is whether $w_B$, the input price in the market with a lower value of time, is higher than $w_A$, the price in the market with higher value of time. In absence of congestion effects and, therefore, of interrelation between markets, the input price is higher in the market with the higher inverse demand intercept (see e.g. Arya and Mittendorf, 2010). This is because with linear demands, the intercept determines the elasticity of the derived demand faced by the input supplier. The market with the higher inverse demand intercept is the less elastic market under uniform prices and, therefore, the market where the discriminating input price will be higher (i.e. the “strong” market). We seek to understand what is the effect of the congestion externality on this. Let $A_i$ be the intercept of the inverse demand function for market $i$. Assuming that the second-order conditions are satisfied, the following lemma summarizes the condition for $w_B > w_A$ to hold (the proof of this Lemma and all other proofs required in this section are in Appendix 5.B):

$$
A_{i} \text{ sufficient condition is that time valuations are not too distinct in that } v_B/v_A > 7 - 4\sqrt{3} \approx 0.072.
$$
**Chapter 5 Input third-degree price discrimination by congestible facilities**

**Lemma 5.1.** The input price under price discrimination is higher in the market with a lower value of time \((w_B > w_A\) holds) if, and only if,

\[
\frac{A_B}{A_A} > \lambda_1 = \frac{8 \cdot P'_A \cdot P'_B + 5 \cdot v_A \cdot v_B \cdot D'^2 + v^2_B \cdot D'^2 + 2 \cdot v_B \cdot D' \cdot \left[-4P'_A - P'_B\right] - 6 \cdot P'_B \cdot v_A \cdot D'}{8 \cdot P'_A \cdot P'_B + 5 \cdot v_A \cdot v_B \cdot D'^2 + v^2_A \cdot D'^2 + 2 \cdot v_A \cdot D' \cdot \left[-4P_B - P_A\right] - 6 \cdot P'_A \cdot v_B \cdot D'},
\]

with \(\lambda_1 < 1\).

To understand the intuition behind the Lemma, first consider the case where time valuations are the same in both markets \((v_A = v_B)\). In this case, \(\lambda_1 = 1\) holds and the result obtained in absence of interrelation goes through. When cross congestion effects are symmetric the supplier’s incentive to charge a higher price in one market over the other do not change and the input price is higher in the market where the reservation price is higher. Second, consider the case where the reservation price is the same in both markets \((A_B/A_A = 1)\), a case where, in absence of congestion and interrelation, it is optimal for the supplier to set a uniform price because the elasticities of the derived demand are the same. If there were only within-market congestion externalities (i.e. absence of interrelation), as in Adachi (2005), it would also be optimal for the facility to set a uniform price. Adachi (2005) shows in final good markets that the price is higher in the market with higher reservation price because it fully determines which market is less elastic when consumption externalities are linear in the quantity. In the case of input markets, this is also the case as it is straightforward to show that the differences in elasticity of the input demand can be fully explained by differences in the reservation price due to the linear demand and congestion assumption. Thus, the cross congestion effects drive the incentive to set a uniform price because the elasticities of the derived demand do not change and the input price is higher in the market where the reservation price is higher. Consequently, when the reservation price is the same in both markets, it is optimal for the input supplier to set a higher price in the market with low time valuation. Phrased differently, for the input supplier the decreased congestion is more profitable in the market with high time valuations because the increase in willingness to pay is increased. Consequently, when the reservation price is the same in both markets, it is optimal for the input supplier to set a higher price in the market with low time valuation. Phrased differently, for the input supplier the decreased congestion is more profitable in the market with high time valuations because the increase in willingness to pay is higher. Third, consider the case of different inverse demand intercepts and time valuations, where both effects come into play as Lemma 5.1 reveals. A lower demand intercept makes the input demand more elastic as it is normally the case with linear demands (in absence of congestion), which gives incentives to decrease the input price, and a lower value of time gives incentives to increase the price in that market because of cross congestion effects. It is straightforward to show that \(\lambda_1\) decreases as the ratio \(v_B/v_A\) is lower, so that the more asymmetric the congestion effects are, the stronger the incentives to raise the price in market \(B\). This is why \(\lambda_1 < 1\) and even when the reservation price is larger in market \(A\), the input price can be higher in market \(B\).

In the general case, the interplay between the relative size of the inverse demand slopes, the time valuations and the inverse demand slopes determines which market faces a higher input price. Interestingly, it is also straightforward to show that \(\lambda_1 > v_B/v_A\) holds regardless of the relative size of the inverse demand slopes. Therefore, Lemma 5.1 implies that it is more likely that the input price is higher in market \(B\) if the asymmetry of inverse demand intercepts is lower than the asymmetry of time valuations. If they are similar or \(A_B/A_A\) is less than the ratio of time valuations, then the input price will be...
higher in market $A$. For the congestion effects to overturn the incentives provided by different demand intercepts (elasticities in absence of congestion), the difference between time valuations must be higher than the difference between demand intercepts.

### 5.3.2 Uniform pricing

Under a uniform pricing regime, the profit-maximizing facility maximizes:

$$\Pi_U = w \cdot [q_A(w, w) + q_B(w, w)] \quad (5.8)$$

and the first-order condition leads to the following pricing rule (again prices are not informative and are in Appendix 5.B):

$$w = 2 \cdot q_A \left[ -P'_A + v_A \cdot D' \right] \cdot \frac{-P'_B + v_B \cdot D'}{-P'_A - P'_B + \bar{v} \cdot D'}$$

$$+ 2 \cdot q_B \left[ -P'_B + v_B \cdot D' \right] \cdot \frac{-P'_A + v_A \cdot D'}{-P'_A - P'_B + \bar{v} \cdot D'}$$

$$- \frac{(q_A + q_B)}{2} \cdot \frac{[v_A \cdot D'] \cdot [v_B \cdot D']}{-P'_A - P'_B + \bar{v} \cdot D'} \quad (5.9)$$

where $\bar{v} = (v_A + v_B)/2$ is the average value of time.

The pricing rule in Eq. (5.9) includes a weighted sum of the supplier’s markups and a negative term that is also a weighted sum of the marginal congestion cost that is external to each firm. It is straightforward to show that the uniform price in Eq. (5.9) is not an weighted average of the differentiated prices in Eqs. (5.6) and (5.7). What causes this result is the demand interdependency through congestion. As Czerny and Zhang (2015) show in final good markets, the uniform price can be lower than both discriminatory prices in presence of congestion externalities. We obtain a similar result in our setting of input price discrimination, so the uniform price is not necessarily an average of the differentiated prices because of demand interrelation. In the following section we study the relationship between uniform and discriminatory prices in detail.

### 5.3.3 The effects of price discrimination on prices and output

To study the effect of price discrimination on input prices and output, we use the price-difference constraint method used by Leontief (1940) and Schmalensee (1981). We assume that the facility maximizes profit subject to the constraint $w_B - w_A \leq t$. This is, the input supplier cannot differentiate prices more than the exogenous amount $t \geq 0$. When $t = 0$, the facility sets the uniform price derived above (Eq. (5.9)). As $t$ gradually increases, the input supplier is gradually allowed to increase the price differentiation until it reaches a point, $t^*$, where it sets the prices $w_A$ and $w_B$ in Eqs. (5.6) and (5.7). The method consists of evaluating the marginal effect of relaxing the constraint on a variable, such as aggregate output. If the sign of the marginal effect does not change in the range $[0, t^*]$, the overall effect of price discrimination on the variable will have the same sign, as long as the unrestricted input provider sets a higher charge in market $B$ ($w_B > w_A$). If the opposite
Chapter 5 Input third-degree price discrimination by congestible facilities

holds, i.e. \( w_B < w_A \), the overall effect of price discrimination will have the opposite sign of the marginal effect, because the price discrimination behavior is approached by making \( t \) negative. All the derivations needed for the results in this section are in Appendix 5.B.

For a given value of \( t \in [0, t^*] \), the facility maximizes:

\[
\Pi = w_A \cdot q_A(w_A, w_A + t) + (w_A + t) \cdot q_B(w_A, w_A + t). \tag{5.10}
\]

Totally differentiating the first-order condition \( \partial \Pi / \partial w_A \), we can obtain the marginal effect on the aggregate output and input prices:

\[
\frac{dQ}{dt} = \left[ v_A - v_B \right] \cdot \frac{D'}{2 \cdot \Omega_1} > 0, \tag{5.11}
\]

\[
\frac{dw_A}{dt} = \frac{4 \cdot P'_A - [3 \cdot v_A - v_B] \cdot D'}{\Omega_2} < 0, \tag{5.12}
\]

\[
\frac{dw_B}{dt} = \frac{[3 \cdot v_B - v_A] \cdot D' - 4 \cdot P'_B}{\Omega_2}, \tag{5.13}
\]

where \( \Omega_1 \) and \( \Omega_2 \) are positive constants. The results that follow from Eqs. (5.11)–(5.13) are summarized in the following proposition.

**Proposition 5.1.** When demands are such that the facility sets a higher input price in the market whose consumers have a lower value of time (i.e. \( w_B > w_A \)), price discrimination:

(i) Increases aggregate output.

(ii) Decreases the input price in the market where time valuations are higher \((A)\).

(iii) Decreases both input prices if time valuations are sufficiently different in that \( v_A - 3 \cdot v_B > -4 \cdot P'_B / D' \); otherwise, it increases the input price in the market where time valuations are lower \((B)\).

When demands are such that \( w_A > w_B \) holds, the effects are reversed and price discrimination decreases output, increases the input price in market \( A \), and it increases both prices when time valuation are sufficiently different.

Our output effect result is an extension of the result in Layson (1998), who shows, in substitute final good markets, that under linear demands the sign of the output effect is determined by the relative magnitude of the gross substitution effect. In our setting outputs are not substitutes nor complements, but the interdependency through congestion generates a similar effect as substitution. An output increase in one market increases the full price of the other market’s consumers by means of increased congestion and therefore it induces an output reduction. As the cross effects are proportional to the time valuations (see Appendix 5.B), it is intuitive that the output change is not zero as long as these effects are not symmetric \((v_A \neq v_B)\). A difference with Layson (1998) is that the relative magnitude of the cross effects is not the only determinant of the sign of the output effect. That is, if \( v_A > v_B \) holds and the relative magnitude of the cross effects is given, output may rise or fall depending on the relative magnitude of the input prices (the sign of \( w_B - w_A \)).
This result on the output effect also extends previous analyses of third-degree price discrimination in presence of congestion externalities. Czerny and Zhang (2015) find that price discrimination by a monopoly airline to two classes of passengers (where, as in our case, demands are only interrelated though congestion) always reduces the aggregate quantity under linear demands. The key difference lies in the properties of the derived demands, which can differ essentially with the final good demands. Czerny and Zhang (2015) assume that demands are such that the (final good) price is higher in the market where time valuations are higher and this implies that price discrimination cannot increase output. Although this seems adequate in their setting as, for example, one can think of business and leisure passengers, it is not necessarily appropriate for input prices. This is because the assumption that there is a strong market where consumers pay a higher price in equilibrium is not necessarily a good proxy for the relative magnitude of the input prices under price discrimination. It is possible that $w_B > w_A$ holds and that the equilibrium downstream price is higher in market $A$.

The effect of price discrimination on prices also extends Layson (1998). He shows, under linear demands, that for prices to move in the same direction when price discrimination is allowed, cross-price effects must be asymmetric and the firm’s marginal costs must be decreasing. In our setting, only asymmetric cross effects are required ($v_A \neq v_B$) as marginal costs are constant. The difference here is that if the (input) price rises in one market, the aggregate quantity does not necessarily decrease because of congestion effects: when the price in one market increases, the full price in the other market may decrease because of the decreased congestion costs. The results also extend Czerny and Zhang (2015) who find that price discrimination cannot reduce both prices, but only increase them. Again the difference is in their assumption on the relative magnitude of final prices.

The intuition of why price discrimination can reduce both prices is similar to the one provided by Layson (1998). In our model, an increase in the input price of one market increases the profitability of the other market, as congestion costs decrease. This increases the consumers’ willingness to pay and therefore the price that can be charged. Under uniform pricing, the marginal profit of the input provider in each market has a different sign. Consider that the marginal profit is negative for market $A$ under uniform pricing (consistent with $w_B > w_A$). If the marginal profit increases slowly towards zero, the decrease in price towards the optimally differentiated $w_A$ will be large. This large decrease may cause a large reduction in the profitability in market $B$, which was positive at uniform prices, and can make it negative at $\{w_A, w\}$. This will therefore cause a reduction also in the price in market $B$. This is what happens when the facility sets a higher input price in the market whose consumers have a lower value of time and time valuations are sufficiently different in that $v_A - 3 \cdot v_B > -4 \cdot P_B'/D'$. A similar explanation works for the case where both prices increase.

Which of the results in Proposition 5.1 is more likely to take place depends on the relation between time valuations and reservation prices. In our model, there are three sources of asymmetry between markets: time valuations, reservation prices and inverse demand slopes. All three are arguably correlated through (average) income: a higher income in market $A$ would explain a higher time valuation, and it would also imply that the reservation price is higher and the demand less sensitive to price changes. The sign of the output effect depends on whether $w_B > w_A$ holds or not. Using Lemma 5.1, we obtain
that output is more likely to increase with price discrimination when the ratio $A_B/A_A$ is greater than the ratio of time valuations $v_B/v_A$ and it will decrease if the asymmetry in reservation prices is similar to or higher than the asymmetry of time valuations. This, naturally, will depend on how the differences in income impacts the time valuations and reservation prices and it is ultimately a matter of empirical investigation. Second, the way in that prices change with price discrimination depend also on how asymmetric the time valuations are. Price discrimination is likely to move prices in the opposite direction when the ratio of time valuations $v_B/v_A$ is not too low (higher than 1/3 is sufficient) and to change prices in the same direction when it is sufficiently low ($v_B/v_A$ at least lower than 1/3). As explained above, this is because when they are sufficiently different ($v_A - 3 \cdot v_B > -4 \cdot P'_B/D'$), the change in profitability in one market due to the change in the input price of the other is large.

The likelihood of $v_A - 3 \cdot v_B > -4 \cdot P'_B/D'$ is somewhat difficult to assess. One way of casting light into its likelihood is by considering that the differences across markets are caused by differences in trip purpose. Koster et al. (2011), Kouwenhoven et al. (2014) and Shires and De Jong (2009) provide empirical evidence that the ratio of time valuations between business and other users in transport markets is not higher than 3. This suggests that $v_A - 3 \cdot v_B > -4 \cdot P'_B/D'$ is a rather stringent condition when differences between markets are caused by differences in the proportion of business and other types of travelers. In that case it is more likely that input price discrimination increases the price in one market and decreases the price in the other. Which market faces the high price depends on the relation between time valuations and reservation prices as discussed earlier. If the ratio of demand intercepts is similar as the ratio of time valuations, for example, because income affects reservation prices and time valuations in a similar way, then price discrimination will increase the price in the market with high time valuations (market $A$). Only when the ratio $A_B/A_A$ is greater than the ratio of time valuations $v_B/v_A$ price discrimination can increase the input price in the low income market (market $B$).

### 5.3.4 Welfare analysis

A full characterization of the marginal welfare effect would be tedious in our case. First, unlike the case of final good markets, under the uniform pricing regime there is, in general, a misallocation of output between markets. This is because downstream firms charge a markup related to demand characteristics and time valuations, so that when the input price is uniform, the marginal willingness to pay is, generally, not the same in each market. To see this, consider the marginal change in total welfare as more discrimination is allowed using the same method as in the previous section:

$$
\frac{dW}{dt} = \frac{dq_A}{dt} \left[ (w_A - w) - q_A \cdot P'_A + q_A \cdot v_A \cdot D' \right] + \frac{dq_B}{dt} \left[ (w_B - w) - q_B \cdot P'_B + q_B \cdot v_B \cdot D' \right] + \frac{dQ}{dt} \cdot \left[ w - (q_A \cdot v_A + q_B \cdot v_B) \cdot D' \right],
$$

(5.14)

where the first two terms in square brackets are the final good prices $(w_i - q_i \cdot P'_i + q_i \cdot v_i \cdot D')$ minus the uniform input price input price set by the facility $(w)$, and the third bracketed term is the difference between the uniform input price and the marginal
external congestion cost. When the input prices are uniform and equal to \( w \), there is still a misallocation effect unless the sum of the demand related markup and the internalized congestion is the same in both markets \((-q_A \cdot P'_A + q_A \cdot v_A \cdot D' = -q_B \cdot P'_B + q_B \cdot v_B \cdot D')\).

Second, as Czerny and Zhang (2015) point out, the presence of congestion externalities gives rise to an effect that works in the opposite direction as the output effect on welfare. Thus, welfare can increase when output is decreased by price discrimination. This section provides a partial characterization of the effect of price discrimination on welfare by deriving sufficient conditions for welfare improvement and deterioration. Rearranging Eq. (5.14), we get:

\[
\frac{dW}{dt} = \frac{dq_A}{dt} \cdot \left[ w_A - \left[ q_A \cdot P'_A + q_B \cdot v_B \cdot D' \right] \right] + \frac{dq_B}{dt} \cdot \left[ w_A + t - \left[ q_B \cdot P'_B + q_A \cdot v_A \cdot D' \right] \right],
\]

(5.15)

where the terms in square brackets multiplying the marginal quantity changes are the difference between the input price set by the facility and the socially optimal input price.

The welfare analysis can be divided in two cases, namely when price discrimination changes both quantities in the same direction (both either rise or fall) and when price discrimination increases the quantity in one market and it decreases it in the other. We first focus in the latter case. Opposite changes in demand due to price discrimination are a consequence of opposite changes in prices. As discussed in Proposition 5.1, this happens when time valuations are not too different and, thus, the effect of a price change in one market on the marginal profitability of the other is not large enough to provide incentives to increase or decrease both input prices. In this case, output increases in the market where the input price decreases and it decreases in the other market. We provide sufficient conditions for welfare improvement when aggregate output increases and for welfare deterioration when aggregate output decreases. As shown in Proposition 5.1, the sign of the output effect depends on whether \( w_B > w_A \) holds or not, which depends on whether \( A_B/A_A > \lambda \) holds or not. First, if demands are such that \( w_B > w_A \) \((A_B/A_A > \lambda_1)\), the aggregate output increases and, therefore, the quantity decrease in market \( B \) is lower than the increase in market \( A \). As a consequence, from Eq. (5.15), if the difference in actual and socially optimal input price is positive in market \( A \) and higher than in market \( B \) for all values of \( t \), then price discrimination necessarily increases welfare. Conversely, if \( w_B < w_A \) holds \((A_B/A_A < \lambda_1)\), the aggregate output decreases and the quantity decrease in market \( A \) is higher than the increase in market \( B \). Therefore, if the difference in actual and socially optimal input price is always positive and higher in market \( A \), welfare decreases. The conditions for this are summarized in the following proposition:

Proposition 5.2. When time valuations are similar in that \( v_A - 3 \cdot v_B < -4 \cdot P'_B/D' \), the quantities change in opposite directions with price discrimination and:

(i) Price discrimination increases welfare if:

\[
\lambda_1 < \frac{A_B}{A_A} < \lambda_2 = \frac{12 \cdot P'_A \cdot P'_B + 10 \cdot v_A \cdot v_B \cdot D'^2 + 2 \cdot v_B^2 \cdot D'^2 + 3 \cdot v_B \cdot D'}{12 \cdot P'_A \cdot P'_B + 10 \cdot v_A \cdot v_B \cdot D'^2 + 2 \cdot v_A^2 \cdot D'^2 + 3 \cdot v_A \cdot D'},
\]

where the terms in square brackets multiplying the marginal quantity changes are the difference between the input price set by the facility and the socially optimal input price.
Chapter 5  Input third-degree price discrimination by congestible facilities

(ii) Price discrimination decreases welfare if:

\[
\frac{A_B}{A_A} < \min \{ \lambda_1, \lambda_3 \}, \quad \text{where} \quad \lambda_3 = \left[ \frac{-P_A' + v_A \cdot D'}{-P_B' + v_B \cdot D'} \right] \cdot \frac{-4 \cdot P_B' + v_A \cdot D' + 5 \cdot v_B \cdot D'}{-4 \cdot P_A' + v_B \cdot D' + 5 \cdot v_A \cdot D'}.
\]

Let us first discuss part (i), where \( w_B > w_A \) holds and aggregate quantity increases. The reason why welfare increases is that the benefit in the market \( A \) from a decreased input price and increased quantity is larger than the loss in market \( B \), where the opposite happens. Therefore, it follows that demand in market \( B \) cannot be significantly larger than in market \( A \) for this to hold. This is why an upper bound on \( A_B/A_A \) is needed. It is expected that the market with high time valuations is also the market with low demand price sensitivity, so that \( v_A > v_B \) and \( -P_A' > -P_B' \) hold. In this case, it is straightforward to show that \( \lambda_2 < 1 \). In addition, the interval \( [\lambda_1, \lambda_2] \) is non-empty when the ratio of the inverse demand slopes is less than the ratio of time valuations, i.e. \( v_B/v_A < | -P_B' / A - P_A'/A | \). Thus, price discrimination is likely to increase welfare when time valuations are similar (\( v_B/v_A > 1/3 \) is sufficient) and the price sensitivities as well as the reservation prices are more similar.

The second part of Proposition 5.2 is intuitive: if \( A_B/A_A \) is lower than \( \lambda_1 \), price discrimination increases the price in the market \( A \), which is the market with a higher reservation price and with higher time valuations. This, not surprisingly, is likely to reduce welfare. If \( -P_A' > -P_B' \) holds, which is expected to hold if the difference of time valuations across markets is due to differences in income, \( \lambda_3 > \lambda_1 \) holds and therefore \( A_B/A_A < \lambda_1 \) is a sufficient condition for welfare deterioration. \( \lambda_3 \) is only part of the necessary condition in the case where \( -P_A' < -P_B' \) and it is sufficient for the welfare loss in the market \( A \) to be higher than the gain in \( B \). As a result, if time valuations are similar (a ratio higher than 1/3 is sufficient), the market with high time valuations is also the market with lower demand sensitivity to price changes and the ratio of demand intercepts is similar to or lower than the ratio of time valuations, price discrimination will decrease welfare.

The welfare analysis when price discrimination changes both quantities in the same direction is in Appendix 5.B. We choose not to discuss it here because the conditions that make price discrimination to increase or decrease both quantities are rather stringent and not very informative. We show that both prices moving in the same direction is not sufficient for both quantities to move in the same direction because of congestion effects. A change in demand in one market has an impact on the full price of the other market and, if cross congestion effects are not low, this can overturn the effect of the own input price change. The conditions that make quantities to either rise or fall involve an upper and lower bound on the ratio of time valuations and also a restriction on the relationship between time valuations and demand slopes. Moreover, the time valuation in market \( A \) has to be more than 5 times larger than in market \( B \), which, as we argue above, seems to be unrealistic in transport markets. Nevertheless, when price discrimination increases the quantity in both markets it increases consumer surplus in both markets and total welfare. The reverse may also happen and price discrimination can decrease both quantities, decrease welfare and consumer surplus.

The results of this section show benefits from input price discrimination in the presence of negative consumption externalities and that price discrimination can increase consumer surplus. Importantly, the benefits are found in a setting where, in absence of externalities,
price discrimination yields lower social welfare. In addition, the conditions for welfare improvement depend strongly on the absolute and relative value of the congestion effects. This suggests that the efficiency of a pricing policy can differ with the level of congestion of the facility even if everything else is invariant (e.g. through different capacity of the facility).

A natural expectation when the differences in time valuations arise from differences in income across markets is that the market with high time valuations (A) also exhibits lower demand sensitivity to price changes ($-P'_A > -P'_B$) and a higher reservation price ($A_A > A_B$). In this case, it follows from Propositions 5.1 and 5.2 that price discrimination is more likely to decrease welfare when the asymmetry of the reservation prices is similar to or higher than the asymmetry in time valuations (i.e. when $A_B/A_A \leq v_B/v_A$). This is because under linear demands and congestion the market with a higher demand intercept is the less elastic market. It also follows that for price discrimination to increase welfare, the ratio of reservation prices $A_B/A_A$ needs to be higher than the ratio of time valuations.

In addition, when time valuations are similar in that $v_B/v_A \geq 1/3$ holds, price discrimination is more likely to increase welfare when the ratio $A_B/A_A$ is, besides being larger than $v_B/v_A$, not too high. For example, when $1/3 < v_B/v_A < 1/2$, price discrimination decreases welfare when $A_B/A_A < 2/3$ and it can increase welfare if $2/3 < A_B/A_A < 1$.

In the following section we analyze price discrimination by a public facility with the aim of comparing the welfare results and shed light on when a broad ban on price discrimination is desirable.

5.4 Public facility

We now study a public facility that maximizes domestic welfare. If the facility were maximizing total welfare, allowing price discrimination would always be optimal and the analysis would be trivial. We introduce a source of divergence from total welfare maximization, namely that consumers and firms may be foreign. Among the many possible domestic-foreign structures we consider the case where market $A$ is fully domestic (passengers and firm $A$ are domestic) and the firm $B$ together with a fraction of the passengers in market $B$ are foreign. The assumption that the market with higher time valuations is the domestic market is, we believe, a realistic assumption if the differences in income across markets are a consequence of differences in trip purpose, as business travel is more frequent in domestic destinations than in international travel. For example, in 2012, the share of business trips was 20%, 30% and 190% higher in domestic destinations than in international travel at London City (LCY), London Heathrow (LHR) and Manchester (MAN) airports respectively (CAA, 2012). In 2011 in Los Angeles International Airport (LAX), the share of business trips was 90% higher in U.S. destinations than in international destinations (Unison Consulting, 2011).

56 Our assumption may be less realistic for air transportation in high income countries with small domestic markets, such as the Netherlands or Switzerland. In those cases, our model may be representative of other transportation markets where congestible facilities provide an input to downstream firms, such as rail transportation.
therefore, the public facility maximizes the sum of its profit, firm A’s profit, the consumer surplus in market A and one half of the consumer surplus in market B:

\[ W_D = \left[ \int_0^{q_A} P_A(x) dx - v_A \cdot q_A \cdot D(Q) \right] + \frac{1}{2} \cdot \left[ \int_0^{q_B} P_B(x) dx - q_B \cdot P_B(q_B) \right] + [w_B \cdot q_B]. \]

(5.16)

where the first term in square brackets is total welfare in market A (the sum of the consumer surplus, firm A’s profit and airport revenues from market A), the second term in square brackets is the consumer surplus in market B and the third term is the airport’s revenue from market B.

The incentive to price discriminate is to capture part of the foreign firm’s profit and stimulate domestic production. The model can easily be extended to cases where a different share of consumer surplus is taken into account by the facility or to cases where there are foreign passengers in both markets, but results do not change in any significant way. What matters is that there is a clear incentive to reduce the price in one market in detriment of the other, and not so much which is the mechanism that provides this incentive.

5.4.1 Price discrimination

The first-order conditions of maximizing \( W_D \) with respect to both input prices lead to the input prices \( w_A \) and \( w_B \) (see Appendix 5.C for the prices and all derivations of the results in this section). Here, as in the previous section, we present the pricing rules:

\[ w_A = q_A \cdot P'_A + q_B \cdot v_B \cdot D', \]
\[ w_B = 2 \cdot q_B \cdot \left[ -\frac{3}{4} \cdot P'_B + v_B \cdot D' \right] + q_A \cdot v_A \cdot D'. \]

(5.17)

(5.18)

The input price for the domestic firm is a subsidy equal to the downstream markup \( (q_A \cdot P'_A < 0) \) and the marginal congestion cost that is not internalized by firm A \( (q_B \cdot v_B \cdot D') \).\(^{57}\)

This is the first-best pricing rule, as it makes the final price in the market equal to the marginal social cost (see Eq. (5.4)). The price in the foreign market is the sum of a market power markup and the marginal congestion cost that is not being internalized. The public facility does not subsidize the foreign firm, but the markup is lower than in the private case, as the consumer surplus in this market is partially taken into account because a fraction of the consumers are domestic.

From comparing the pricing rules above, it follows that \( w_B > w_A \) always holds in this case, a result of the assumed domestic-foreign structure. This captures the usual argument to enforce uniform pricing by a public supplier that it protects consumers of foreign markets. In addition, \( w_B \) is always positive and the sign of \( w_A \) is ambiguous and depends on whether the inefficiency due to downstream market power is larger or smaller than the inefficiency due to the congestion externality.

\(^{57}\)The facility charges the externality imposed on the foreign market because it is profit maximizing to do so (see the pricing rule of the private facility in Eq. (5.6)) and not because of consumer surplus considerations.
5.4.2 Uniform pricing

When the facility is restricted to charge the same input price to both firms, we obtain the following pricing rule:

\[
\begin{align*}
\frac{dw}{dt} &= q_A \cdot P'_A \cdot \frac{2 \cdot [P'_B + v_B \cdot D'] - v_A \cdot D'}{2 \cdot [(P'_A - P'_B) + (v_A + v_B) \cdot D']} \\
&\quad - q_B \cdot P'_B \cdot \frac{3 \cdot [P'_A + v_A \cdot D'] + v_B \cdot D'}{2 \cdot [(P'_A - P'_B) + (v_A + v_B) \cdot D']} \\
&\quad + q_A \cdot v_A \cdot D' \cdot \frac{2 \cdot [P'_A + v_A \cdot D'] - v_B \cdot D'}{2 \cdot [(P'_A - P'_B) + (v_A + v_B) \cdot D']} \\
&\quad + q_B \cdot v_B \cdot D' \cdot \frac{4 \cdot [P'_A + v_A \cdot D'] - v_A \cdot D'}{2 \cdot [(P'_A - P'_B) + (v_A + v_B) \cdot D']}. \tag{5.19}
\end{align*}
\]

The pricing rule in Eq. (5.19) includes a weighted sum of the subsidy for firm A and the markup for firm B present in the discriminating prices. It also includes a weighted sum of the market-specific marginal congestion cost that are also part of the differentiated input prices. We elaborate in the following section on the relation between the uniform and the discriminating input prices and show that the uniform price is a weighted average of the discriminatory prices.

5.4.3 The effects of price discrimination on prices and output

Using the same price difference constraint method as in Section 5.3, we obtain the following results regarding the marginal effect of price discrimination on the aggregate output and on input prices:

\[
\begin{align*}
\frac{dQ}{dt} &= \frac{2 \cdot [P'_A + v_A] \cdot [-5 \cdot P'_B + 4 \cdot v_B - v_A] + v_B \cdot [3 \cdot P'_B - 2 \cdot v_B]}{\Omega_3}, \tag{5.20} \\
\frac{dw_A}{dt} &< 0, \tag{5.21} \\
\frac{dw_B}{dt} &> 0, \tag{5.22}
\end{align*}
\]

where \( \Omega_3 \) is a positive constant. The results that follow from Eqs. (5.20)–(5.22) are summarized in the following proposition.

**Proposition 5.3.** Price discrimination by a public facility:

(i) Increases the aggregate output if time valuations are sufficiently similar in that \( v_A - 3 \cdot v_B < -4 \cdot P'_B/D' \).

(ii) Decreases the input price in the market served by the domestic firm.

(iii) Increases the input price in the market served by the foreign firm.
This is intuitive, when the facility is allowed to price discriminate it reduces the price in the market served by the domestic firm and raises the price to the foreign firm to capture part of its profit. When the condition (i) in Proposition 5.3 holds, the output increase in the market served by the domestic firm is larger than the decrease in the market served by the foreign firm.

5.4.4 Welfare analysis

Unlike in the case of a private facility, we can analyze the welfare effect directly as opposed to using the price difference constraint method used in Section 5.3. Recall that in this section we look at how total welfare changes when a facility that maximizes domestic welfare is allowed to differentiate prices. The main result of the analysis is summarized in the following proposition.

**Proposition 5.4.** Price discrimination by a public facility increases total welfare if, and only if, \( \frac{A_B}{A_A} < \lambda_4 \), and it decreases total welfare when \( \frac{A_B}{A_A} > \lambda_4 \).

Where \( \lambda_4 \) is a fraction whose numerator and denominator are a function of the demand sensitivity parameters \((P'_A, P'_B)\) and of the congestion effects \((v_A \cdot D'_A, v_B \cdot D'_B)\) in a similar way as \( \lambda_1 \) and \( \lambda_2 \). However, both the numerator as well as the denominator of \( \lambda_4 \) are polynomials of degree 7, so we omit the expression here (see Appendix 5.C).

From a total welfare standpoint, price discrimination leads to a welfare loss in the market served by the foreign firm \((B)\), as the price moves away from the marginal social cost, and to a welfare gain in the market served by the domestic firm \((A)\), as the price moves towards marginal social cost. To obtain intuition for the result in Proposition 5.4 consider the case where there are no congestion effects \((v_A = v_B = 0)\). In this case, \( \lambda_4 \) is a function only of the demand slopes and if the inverse demand is steeper in market \(A\) (i.e. \(-P'_A > -P'_B\)) \( \lambda_4 > 1 \) holds. Therefore, in absence of congestion, if the foreign market is more elastic at uniform prices \((A_A > A_B)\) and it also has a higher sensitivity to price changes, total welfare increases with price discrimination. This result is natural, price discrimination raises the price in market \(B\), which is the more elastic and the more price sensitive, so the welfare losses due to the double marginalization are limited compared to the gains of pricing the domestic market at marginal social cost. In the general case, congestion effects come into play and patterns are complex. Increased demand can have negative effects and the cross-effects that resemble substitution may change the conclusions. Nevertheless, the result in Proposition 5.4 that there is an upper bound for \(A_B/A_A\) for total welfare improvement is intuitive. Price discrimination is more likely to increase total welfare when the foreign market –where the input price is raised– is relatively more elastic and not too large. For example, when \(v_B/v_A = 1/3\), the lowest ratio of time valuations that ensures that output increases with price discrimination, \( \lambda_4 > 1/3 \) holds. As a result, price discrimination increases total welfare when the ratio of the demand intercepts is the same as or lower than the ratio of time valuations. When the time valuations are not more asymmetric than the reservation prices, the congestion effects do not overturn the results obtained in absence of congestion. Importantly, this is in sharp contrast with the results for a private facility where price discrimination is
5.5 Comparison of the welfare effect under private and public ownership

likely to reduce total welfare when the two ratios \((v_B/v_A \text{ and } A_B/A_A)\) are similar. In the following section we argue that \(\lambda_4 > v_B/v_A\) is likely to hold more generally, so the conclusion that price discrimination by a public facility increases total welfare when the ratio of the reservation prices is similar to the ratio of time valuations is not restricted to the particular case of \(v_B/v_A = 1/3\).

In the following section we also compare the effect of allowing price discrimination on total welfare under both ownership forms in more detail and study whether a broad ban, that covers facilities with different ownership forms, is desirable.

5.5 Comparison of the welfare effect under private and public ownership

To compare the welfare effect of price discrimination we focus on what we believe is the most realistic setting: a case in which, in market \(A\), the time valuation is higher, the demand sensitivity to price changes is lower and the reservation price is higher than in market \(B\). This is a natural expectation if the differences across markets are caused by differences in income and average income is higher in market \(A\). Consequently, throughout this section we assume that \(v_A > v_B, -P'_A > -P'_B\) and \(A_A > A_B\) hold.

We also limit the comparison to the case where time valuations are similar in that \(v_B/v_A \geq 1/3\) holds, which ensures that price discrimination by a private facility changes prices in opposite directions (see Proposition 5.1). We do not compare the welfare effect of price discrimination when price discrimination by a private facility either increases or decreases the output in each market, because the sufficient conditions for the quantities to move in the same direction are too stringent for the case of a public facility. That is, the parameter region where each firm’s output is positive in equilibrium under price discrimination by a public facility is very limited. Moreover, estimations of time valuations suggest that the condition \(v_B/v_A \geq 1/3\) is realistic. Koster et al. (2011) and Kouwenhoven et al. (2014) estimate the value of access time, the value of schedule delay and the value of travel time savings for business and other travel purposes in Dutch air transport passengers. They find that the ratio between other purposes and business time valuations is higher than 0.5 in all cases. This is a lower bound on the ratio of time valuation between markets, if differences between markets are a consequence of different composition of business and other travelers. A meta-analysis covering 30 countries and 77 studies that estimate values of travel time savings in different modes by Shires and De Jong (2009) find that in average, the ratio between time valuations of commuting travelers and business travelers is 0.4. They also report that the ratio between other purposes and commute is on average 0.84, which implies that the lowest ratio is 0.336 (between non commuting and business travelers). This evidence also supports the relevance of the case we study in this section.

When \(v_B/v_A \geq 1/3\) holds, price discrimination by a public facility increases aggregate welfare when \(\lambda_4 > 1\) holds and price discrimination always increases total welfare when the inverse demand intercept in the market served by the domestic firm is larger than or of the same size as the one served by the foreign firm.

\(^{58}\)The results in Proposition 5.4 are also valid when the market served by the foreign firm is also the one where consumers have higher time valuations \((v_B > v_A)\). In this case, \(\lambda_4 > 1\) holds and price discrimination always increases total welfare when the inverse demand intercept in the market served by the domestic firm is larger than or of the same size as the one served by the foreign firm.
output, increases the input price in market $B$ and decreases the input price in market $A$. Price discrimination by a private facility also increases aggregate output, raises the price in market $B$ and decreases the price in market $A$ if $w_B > w_A$ holds ($A_B/A_A > \lambda_1$). Conversely, if $w_B < w_A$ holds ($A_B/A_A < \lambda_1$) it decreases aggregate output, the price in market $B$ falls and the price in market $A$ increases. The relevant comparison is between Propositions 5.2 and 5.4 and the main results are summarized in the following proposition (the proof follows directly from the propositions):

**Proposition 5.5.** When time valuations are sufficiently similar ($v_B/v_A \geq 1/3$), the market with higher time valuations is also the market with lower demand sensitivity to price changes and it is the domestic market:

(i) A ban on price discrimination is desirable only for a private facility if $\frac{A_B}{A_A} < \min[\lambda_1, \lambda_4]$

(ii) A broad ban on price discrimination is desirable if $\lambda_4 < \frac{A_B}{A_A} < \lambda_1$

(iii) A ban on price discrimination is desirable only for a public facility if $\max[\lambda_1, \lambda_4] < \frac{A_B}{A_A} < \lambda_2$

This summary allows for shedding light on the desirability of a broad ban on price discrimination. First, if the reservation price in the domestic market is significantly larger than the reservation price in the foreign market ($A_A >> A_B$), a ban that covers both ownership forms may not be desirable. In this case it is socially optimal to have a public facility differentiating prices as the potential benefits from increased domestic production in market $A$ are large compared to the losses in market $B$. This is because $A_A >> A_B$ implies that market $B$ is much more elastic at uniform prices and the welfare losses are limited because the markup is limited.

Second, when time valuations are not too close to each other, i.e. $1/3 < v_B/v_A \leq 4/5$, and the asymmetry in time valuations, in reservation prices and in demand sensitivity to price changes is similar, a broad ban on price discrimination is not desirable. This is because $\lambda_1 > v_B/v_A$ holds regardless of other parameters and when $1/3 < v_B/v_A < 4/5$ and $-P'_B/-P'_A > 1/10$ hold, $\lambda_4 > v_B/v_A$ holds. Therefore, whenever the asymmetry in reservation prices is similar to the asymmetry in time valuations ($A_B/A_A \approx v_B/v_A$) or higher ($A_B/A_A < v_B/v_A$), allowing a public facility to differentiate prices raises total welfare and a broad ban cannot be the welfare maximizing pricing policy. The intuition is similar as in the previous case. In absence of congestion $A_B/A_A < 1$ and $-P'_A > -P'_B$ are sufficient for price discrimination by a public supplier to be welfare improving. As explained in the previous section, this is because the higher price sensitivity and the higher elasticity at uniform prices of market $B$ limit the markup and welfare losses. As argued in Section 5.4, congestion effects may overturn this. However, if the time valuations are as symmetric or more symmetric than the reservation prices, it is less likely that the welfare effect is overturned. This is why when $A_B/A_A \leq v_B/v_A$, it is efficient to allow price discrimination by a domestic welfare maximizing facility.

Third, the results above suggest that when time valuations are not too similar (i.e. $1/3 < v_B/v_A \leq 4/5$) a broad ban may be desirable if the asymmetry in reservation prices is lower than the asymmetry in time valuations and if the reservation prices are not too
close to each other. This is because \( \lambda_4 < A_B/A_A < \lambda_1 \), the condition (ii) of Proposition 5.5, is needed, but \( \lambda_1 > v_B/v_A \) and \( 1 > \lambda_1 > v_B/v_A \) hold. However, \( \lambda_4 < \lambda_1 \) does not hold globally. In the limit where \( v_A = v_B = 0 \) and when \( v_A \rightarrow \infty \) it does not hold so a broad ban on price discrimination may not be desirable, but it may hold for intermediate values of \( v_A \). We use numerical examples below to show that this may occur.

Fourth, when time valuations are such that \( 4/5 < v_B/v_A < 1 \) holds, numerical results show that the intuition provided above also holds as long as the asymmetry in inverse demand slopes is not significantly higher than the asymmetry in time valuations. That is, when the ratio of time valuations, inverse demand slopes and reservation prices are similar, a ban on price discrimination is not desirable for both ownership forms also in the case where \( 4/5 < v_B/v_A < 1 \). For example, if \( v_B/v_A = 0.9 \), \( \lambda_4 > v_B/v_A \) holds for all values of \( -P'_B/ -P'_A \) higher than 0.22. In this case, \( A_B/A_A \leq 0.9 \) is sufficient for ban on price discrimination not to be welfare enhancing for both ownership forms.\(^{59}\) Again, if the reservations prices are more similar than time valuations, a broad ban may be desirable.

The reason is the same as for the case \( 1/3 < v_B/v_A \leq 4/5 \). Price discrimination by a public supplier is likely to increase welfare when cross congestion effects are not more asymmetric than the reservation prices and inverse demand slopes.

Finally, from the sufficient conditions derived in Sections 5.3 and 5.4 we cannot assess the desirability of a broad ban when the reservations prices are very similar across markets (i.e. when \( \lambda_1 < A_B/A_A \) and \( \lambda_2 < A_B/A_A \leq 1 \)). Nevertheless, as \( \lambda_1 > v_B/v_A \) holds, the main conclusion that under similar asymmetry across markets a broad ban may not be desirable is not affected by this lack of sufficient conditions for the welfare change.

We complement the results of Proposition 5.5 and the intuition provided with numerical examples. The numerical analysis in Figure 5.1 shows the values of \( \lambda_1 \), \( \lambda_2 \) and \( \lambda_4 \) for different parameter values. In each panel the values of the inverse demand slopes \( -P'_B \) and \( -P'_A \) are fixed and two values of time valuation ratios \( (v_B/v_A) \) are studied. What varies in each figure is the value of time in market \( A \), \( v_A \). The cases where a broad ban on price discrimination is desirable for both ownership forms are highlighted in gray (the condition (ii) in Proposition 5.5 holds). In the white areas, a ban on input price discrimination fails to be efficient for, at least, one ownership form. The numerical analysis reinforces the discussion above that a ratio of reservation prices \( A_B/A_A \) equal to or lower than the ratio of time valuations is a sufficient condition for a broad ban not to be welfare enhancing for both ownership forms. It also shows that \( A_B/A_A \) can be higher than \( v_B/v_A \) and the result still holds. For the studied parameter range, when the ratio of time valuations equals 0.5 a broad ban may not be desirable when the ratio of reservation prices is lower than 0.8. When \( v_B/v_A = 0.85 \), \( A_B/A_A < 0.91 \) ensures the inefficiency of a ban on price discrimination for at least one ownership form. The comparison between Figures 5.1a and 5.1b also confirms that condition (ii) of Proposition 5.5 (\( \lambda_4 < A_B/A_A < \lambda_1 \)) does not hold globally. In our numerical example, a broad ban can be desirable for the case where the ratio of inverse demand slopes equals 1/3 and not when it is 2/3. This suggests that the feasibility of the condition is more closely related with the absolute value of \( -P'_B/ -P'_A \) rather than its relative value with respect to \( v_B/v_A \), as in Figure 5.1b \( v_B/v_A > -P'_B/ -P'_A \)

\(^{59}\)The values of \( -P'_B/ -P'_A \) that ensure that \( \lambda_4 > v_B/v_A \) hold are 0.13, 0.22 and 0.39 for \( v_B/v_A \) equal to 0.85, 0.9 and 0.95 respectively.
Chapter 5 Input third-degree price discrimination by congestible facilities

holds in one case and \( v_B/v_A < -P_B' / -P_A' \) in the other and in neither of them the condition holds. The results in Figure 5.1a also confirm that for relatively low and high values of \( v_A \) a ban cannot be desirable for both types of facilities and it shows that for intermediate values it may be desirable when reservation prices are similar but not too close to each other. In our example, this happens for values of \( A_B/A_A \) between 0.8 and 0.9 when the ratio of time valuations is 1/2 and for values of \( A_B/A_A \) between 0.91 and 0.97 when the ratio of time valuations is 0.85.

Figure 5.1: Sufficient conditions for welfare improvement and deterioration under price discrimination. In the gray areas an ban on input price discrimination is desirable for both ownership forms. In the white areas a ban on input price discrimination fails to be efficient for at least one ownership form. Parameter values: \( P_A' = -1, P_B' = -1/3, D' = 1 \) in Figure 5.1a and \( P_A' = -1, P_B' = -2/3, D' = 1 \) in Figure 5.1b.

Note that for the comparisons of this section we have used sufficient conditions for welfare improvement and deterioration under price discrimination by a private facility instead of comparing the actual effect. Therefore, the regions in which a ban on price discrimination is the socially optimal policy for both ownership forms may be larger than the shaded regions presented in this section.

5.6 Robustness: downstream first-degree price discrimination

In this section we analyze the robustness of the conclusions drawn in the previous section from comparing the effect on total welfare of third-degree price discrimination by a public and a private facility. We study how the main results change when downstream firms apply first-degree price discrimination. This is a theoretical extreme that is useful also to study a situation where there is no downstream inefficiency due to market power and
works as a proxy for a perfectly competitive downstream market.

Downstream firms that perfectly discriminate consumers set a unit price equal to the marginal cost, which is the input price $w_i$ plus the marginal congestion cost that is internal to the firm $q_i \cdot v_i \cdot D'(Q)$, and ask for a premium equal to the surplus of each individual (their willingness to pay net of the experienced delays). This changes the derived demands faced by the input provider which are now a result from the following pricing rules:

$$P_i(q_i) - v_i \cdot D(Q) = w_i + q_i \cdot v_i \cdot D'(Q). \quad (5.23)$$

Following the same methodology as in Sections 5.3 and 5.4, it is possible to derive similar sufficient conditions for welfare improvement and deterioration under third-degree input price discrimination for both ownership forms. The aim of this extension is to analyze how the results in Proposition 5.5 change. Let $\lambda_i'$ be the analogous boundary to $\lambda_i$ derived in Sections 5.3 and 5.4. Proposition 5.5 can be restated in the following way:

**Proposition 5.6.** When downstream firms can perfectly discriminate consumers, time valuations are sufficiently similar ($v_B/v_A \geq 1/3$), the market with higher time valuations is also the market with lower demand sensitivity to price changes and it is the domestic market:

(i) A ban on price discrimination is desirable only for a private facility if $\frac{A_B}{A_A} < \min \{\lambda_1', \lambda_3', \lambda_4'\}$

(ii) A broad ban on price discrimination is desirable if $\lambda_4' < \frac{A_B}{A_A} < \min[\lambda_1', \lambda_3']$

(iii) A ban on price discrimination is desirable only for a public facility if $\max[\lambda_1', \lambda_4'] < \frac{A_B}{A_A} < \lambda_2'$

A difference with respect to Proposition 5.5 is the presence of $\lambda_3'$. When price discrimination by a private facility decreases aggregate output, $\lambda_3'$ is the upper bound for $A_B/A_A$ such that the difference in actual and socially optimal input price is higher in market $A$ than in market $B$ under input uniform pricing. This is a sufficient condition for the loss due to the output contraction in market $A$ to be larger than the benefit from increased production in market $B$. When demand in market $A$ is less price-sensitive than in market $B$ and downstream firms do not price discriminate, the resulting markups ensure that the condition is satisfied and $\lambda_3$ becomes irrelevant. As under downstream perfect price discrimination the unit price is the marginal cost, the condition $A_B/A_A < \lambda_3'$ is needed again. Phrased differently, in absence of downstream markups related to the price sensitivity, the difference between actual and socially optimal input price can be lower in market $A$ than in $B$ under input uniform pricing.

One of our main results is that if the asymmetry in reservation prices, the asymmetry in time valuations and the asymmetry in demand sensitivity to price changes are similar, a broad ban on price discrimination may not be desirable because it is optimal to allow a public facility to price discriminate firms. Under downstream first-degree price discrimination this is also the case if, in addition, the congestion effects are not too low.
This is because \( \lambda'_4 \geq \frac{v_B}{v_A} \) does not hold globally.\(^6^{60}\) To see why, first consider that there is no congestion (i.e. \( v_A = v_B = 0 \)). As there is no downstream market power inefficiency due to the perfect discrimination, in absence of negative consumption externalities, the socially optimal input prices are equal to zero. It can be shown that in this case, price discrimination by a public facility is always welfare decreasing (\( \lambda_4 = 0 \) when \( v_A = v_B = 0 \)), something that does not hold when there is downstream inefficiency due to market power exertion. In the other extreme, in the limit where \( v_A \) goes to infinity, \( \lambda'_4 \) is higher than the ratio of time valuations \( v_B/v_A \) and our main result holds under downstream perfect price discrimination. This implies that if congestion effects are sufficiently high, \( \lambda'_4 > \frac{v_B}{v_A} \) holds and the main welfare result also holds. Therefore, when the asymmetry in reservation prices, the asymmetry in time valuations and the asymmetry in demand sensitivity to price changes are similar across markets and downstream firms perfectly price discriminate, there is always a level of congestion that makes a ban on price discrimination inefficient for at least one ownership form.

The results of this section show that allowing input providers to price discriminate can increase total welfare even in the extreme case of downstream perfect price discrimination. Therefore, the benefits from input price discrimination do not rely on the presence of downstream market power inefficiencies and it can be expected that for imperfect downstream price discrimination the results are closer to those in the previous sections.

To conclude this section we briefly analyze two numerical examples. Figures 5.2a and 5.2b summarize the sufficient conditions for welfare improvement and deterioration when price discrimination is allowed for both ownership forms. We set the parameter values to the values used for the numerical analysis in Figure 5.1b, where the sufficient condition for a broad ban to be desirable was never satisfied. The gray areas display the parameter regions in which a ban on price discrimination is socially optimal for a private as well as for a public facility. Both figures show that, as discussed above, a broad ban on price discrimination can be the optimal policy if congestion effects are not too high. This suggest that the absence of downstream market power inefficiencies enhances the performance of a ban on price discrimination. Figures 5.2a and 5.2b also reveal that when the asymmetry in reservation prices is similar to the asymmetry in time valuations (\( A_B/A_A = 0.5 \) in Figure 5.2a and \( A_B/A_A = 0.85 \) in Figure 5.2b), the extent to which a broad ban on price discrimination may be undesirable is large for the studied parameterization. Finally, the examples suggest that a broad ban is more likely to be desirable, just as in the previous analysis, when the asymmetry of reservation prices is lower than the asymmetry of time valuations (\( A_B/A_A > v_B/v_A \)).

5.7 Conclusions

This chapter has shown how the presence of congestion externalities influences the effects of input third-degree price discrimination. Our framework considers an upstream

\(^6^{60}\) To see why, first consider that there is no congestion (i.e. \( v_A = v_B = 0 \)). As there is no downstream market power inefficiency due to the perfect discrimination, in absence of negative consumption externalities, the socially optimal input prices are equal to zero. It can be shown that in this case, price discrimination by a public facility is always welfare decreasing (\( \lambda_4 = 0 \) when \( v_A = v_B = 0 \)), something that does not hold when there is downstream inefficiency due to market power exertion. In the other extreme, in the limit where \( v_A \) goes to infinity, \( \lambda'_4 \) is higher than the ratio of time valuations \( v_B/v_A \) and our main result holds under downstream perfect price discrimination. This implies that if congestion effects are sufficiently high, \( \lambda'_4 > \frac{v_B}{v_A} \) holds and the main welfare result also holds. Therefore, when the asymmetry in reservation prices, the asymmetry in time valuations and the asymmetry in demand sensitivity to price changes are similar across markets and downstream firms perfectly price discriminate, there is always a level of congestion that makes a ban on price discrimination inefficient for at least one ownership form.

The results of this section show that allowing input providers to price discriminate can increase total welfare even in the extreme case of downstream perfect price discrimination. Therefore, the benefits from input price discrimination do not rely on the presence of downstream market power inefficiencies and it can be expected that for imperfect downstream price discrimination the results are closer to those in the previous sections.

To conclude this section we briefly analyze two numerical examples. Figures 5.2a and 5.2b summarize the sufficient conditions for welfare improvement and deterioration when price discrimination is allowed for both ownership forms. We set the parameter values to the values used for the numerical analysis in Figure 5.1b, where the sufficient condition for a broad ban to be desirable was never satisfied. The gray areas display the parameter regions in which a ban on price discrimination is socially optimal for a private as well as for a public facility. Both figures show that, as discussed above, a broad ban on price discrimination can be the optimal policy if congestion effects are not too high. This suggest that the absence of downstream market power inefficiencies enhances the performance of a ban on price discrimination. Figures 5.2a and 5.2b also reveal that when the asymmetry in reservation prices is similar to the asymmetry in time valuations (\( A_B/A_A = 0.5 \) in Figure 5.2a and \( A_B/A_A = 0.85 \) in Figure 5.2b), the extent to which a broad ban on price discrimination may be undesirable is large for the studied parameterization. Finally, the examples suggest that a broad ban is more likely to be desirable, just as in the previous analysis, when the asymmetry of reservation prices is lower than the asymmetry of time valuations (\( A_B/A_A > v_B/v_A \)).

5.7 Conclusions

This chapter has shown how the presence of congestion externalities influences the effects of input third-degree price discrimination. Our framework considers an upstream
monopoly facility that can be private or public (domestic welfare maximizer) that sells an input to private downstream firms that operate in different markets as monopolists. The presence of downstream within- and cross-market negative externalities makes all demands interrelated in a way that is similar to the case where downstream firms offer substitute products. Using a stylized model with linear demands, we show that aggregate output can increase, all prices can decrease and welfare can increase when discrimination is allowed; although the opposite results are also possible. This is found in a setting in which in absence of congestion, price discrimination by a private input provider leads to welfare deterioration and constant aggregate output. The results of the chapter suggest that the presence of congestion externalities enlarges the extent to which input price discrimination by a private facility is desirable from a welfare standpoint.

We have also analyzed the effects of price discrimination when the supplier maximizes domestic welfare, a common ownership form of transport facilities, and we have compared the welfare effects of price discrimination under private and public ownership. We have characterized the conditions that make the welfare maximizing pricing regime to be different for different ownership forms of the facility. Although the patterns are complex, the main insights are that when the asymmetry in time valuations, reservation prices and price-sensitivity of demand is similar across markets, a ban on price discrimination is not socially optimal for, at least, one type of facility. A broad ban may be desirable when the reservation prices in both markets are significantly more similar than the time valuations. Therefore, the results of this chapter suggest that a broad ban on price discrimination to transport facilities that cover multiple ownership forms, such as the EU Airport Charges directive (2009/12/EC) and the World Trade Organization’s General Agreement on Tar-
iffs and Trade (GATT), may have to be revised. This is especially relevant in the light of the increasing practice of (partially) privatizing transport facilities.

The extent to which the conditions that make the socially optimal pricing regime to be different between public and private ownership hold is an important avenue for empirical future research. The consideration of competition, cost regulations and network effects are also natural extensions for future research. However, as the divergence of the socially optimal policy has been found in the simple framework considered here, it is unlikely to disappear when complexity is added. Finally, the analysis has relied on linear functional forms. Extending the investigation considering other demand and cost functions is a natural and important avenue for future research. Analyzing demand functions with adjusted concavities that make price discrimination by a private facility to decrease output in absence of congestion is of particular relevance to check the robustness of the result that the presence of congestion enhances the performance of input price discrimination.
Appendix 5.A  Glossary of notation

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_i$</td>
<td>Inverse demand intercept for firm $i$ (reservation price)</td>
</tr>
<tr>
<td>$D(\cdot)$</td>
<td>Delay function</td>
</tr>
<tr>
<td>$P_i(\cdot)$</td>
<td>Inverse demand function for firm $i$</td>
</tr>
<tr>
<td>$P_i'$</td>
<td>Inverse demand slope for firm $i$</td>
</tr>
<tr>
<td>$Q$</td>
<td>Aggregate demand over firms and markets</td>
</tr>
<tr>
<td>$q_i$</td>
<td>Quantity of firm $i$</td>
</tr>
<tr>
<td>$v_i$</td>
<td>Value of travel time savings of consumers in market $i$</td>
</tr>
<tr>
<td>$w$</td>
<td>Airport uniform charge</td>
</tr>
<tr>
<td>$w_i$</td>
<td>Airport charge for firm $i$</td>
</tr>
</tbody>
</table>

Table 5.1: Glossary of notation.

Appendix 5.B  Calculations and proofs for Section 5.3

Derived demands

Solving simultaneously the first-order conditions of both downstream firms (see Eq. (5.2)) and denoting $A_i$ the inverse demand intercept in market $i$, we obtain the derived demands:

\[
q_A(w_A, w_B) = \frac{2 \cdot [-P'_B + v_B \cdot D'] \cdot [A_A - w_A] - v_A \cdot D' \cdot [A_B - w_B]}{\Omega_1} \tag{5.24}
\]

\[
q_B(w_A, w_B) = \frac{2 \cdot [-P'_A + v_A \cdot D'] \cdot [A_B - w_B] - v_B \cdot D' \cdot [A_A - w_A]}{\Omega_1} \tag{5.25}
\]

where $\Omega_1 = 2 \cdot [P'_A - v_A \cdot D'] \cdot 2 \cdot [P'_B - v_B \cdot D'] - [v_A \cdot D'] \cdot [v_B \cdot D'] > 0$ \tag{5.26}

And deriving with respect to the input prices, we get:

\[
\frac{\partial q_A}{\partial w_A} = \frac{2 \cdot [P'_B - v_B \cdot D']}{\Omega_1} < 0 \tag{5.27}
\]

\[
\frac{\partial q_A}{\partial w_B} = \frac{v_A \cdot D'}{\Omega_1} > 0 \tag{5.28}
\]

\[
\frac{\partial q_B}{\partial w_A} = \frac{v_B \cdot D'}{\Omega_1} > 0 \tag{5.29}
\]

\[
\frac{\partial q_B}{\partial w_B} = \frac{2 \cdot [P'_A - v_A \cdot D']}{\Omega_1} < 0 \tag{5.30}
\]
**Chapter 5 Input third-degree price discrimination by congestible facilities**

**Input prices**

Solving the first-order conditions for the facility under price discrimination, $\partial \Pi^P D / \partial w_A$ and $\partial \Pi^P D / \partial w_B$, we get:

$$w_A = \frac{[-P'_A + v_A \cdot D'] [8 \cdot A_A \cdot [-P'_B + v_B \cdot D'] - 2 \cdot A_B \cdot [v_A - v_B] \cdot D]}{16 \cdot [-P'_A + v_A \cdot D'] \cdot [-P'_B + v_B \cdot D'] - [v_A \cdot D' + v_B \cdot D]^2}$$

(5.31)

$$w_B = \frac{[-P'_B + v_B \cdot D'] [8 \cdot A_B \cdot [-P'_A + v_A \cdot D'] + 2 \cdot A_A \cdot [v_A - v_B] \cdot D]}{16 \cdot [-P'_A + v_A \cdot D'] \cdot [-P'_B + v_B \cdot D'] - [v_A \cdot D' + v_B \cdot D]^2}$$

(5.32)

Solving $\partial \Pi^U / \partial w$, we obtain:

$$w = \frac{A_A \cdot [-2 \cdot P'_B + v_B \cdot D'] + A_B \cdot [-2 \cdot P'_A + v_A \cdot D']}{4 \cdot [-P'_A - P'_B] + 2 \cdot [v_A + v_B] \cdot D}$$

(5.33)

**Proof of Lemma 5.1**

Using Eqs. (5.31) and (5.32), we get that $w_B - w_A$ equals:

$$\frac{A_B \cdot [8 \cdot P'_A \cdot P'_B + 5 \cdot v_A \cdot v_B \cdot D^2 + v^2_B \cdot D^2 + 2 \cdot v_A \cdot D' \cdot [-4P'_B - P'_A] - 6 \cdot P'_A \cdot v_B \cdot D]}{16 \cdot [-P'_A + v_A \cdot D'] \cdot [-P'_B + v_B \cdot D'] - [v_A \cdot D' + v_B \cdot D]^2}$$

$$- \frac{A_A \cdot [8 \cdot P'_A \cdot P'_B + 5 \cdot v_A \cdot v_B \cdot D^2 + v^2_B \cdot D^2 + 2 \cdot v_B \cdot D' \cdot [-4P'_A - P'_B] - 6 \cdot P'_B \cdot v_A \cdot D]}{16 \cdot [-P'_A + v_A \cdot D'] \cdot [-P'_B + v_B \cdot D'] - [v_A \cdot D' + v_B \cdot D]^2}$$

(5.34)

where the denominator is positive by the second-order conditions of the supplier maximization problem ($v_B/v_A > 7 - 4\sqrt{3} \approx 0.0718$ is sufficient). Therefore, the condition in Lemma 5.1 follows straightforwardly as the terms multiplying $A_A$ and $A_B$ in Eq. (5.34) are positive.

**Effect of price discrimination on output and prices**

To simplify notation, we omit the arguments of the functions and let $\tau$ be the input price in market $A$ and $\tau + t$ the charge in market $B$. For a given $t \in [0, t^*]$, the first-order condition of the supplier’s maximization profit is:

$$\partial \Pi / \partial \tau = [q_A + q_B] + \tau \cdot \left[ \frac{\partial q_A}{\partial \tau} + \frac{\partial q_B}{\partial \tau} \right] + t \cdot \frac{\partial q_B}{\partial \tau}.$$  

(5.35)
This first-order condition defines implicitly $\tau$ as a function of $t$ in the following way:

\[
\frac{d\tau}{dt} = -\frac{\partial^2 \Pi / \partial \tau \partial t}{\partial^2 \Pi / \partial \tau^2} = \frac{\frac{\partial q_A}{\partial w_A} + 2 \cdot \frac{\partial q_B}{\partial w_B} + \frac{\partial q_B}{\partial w_A} + \frac{\partial q_B}{\partial w_B}}{2} [4 \cdot P_A' - [3 \cdot v_A - v_B] \cdot D']/\Omega_1
\]

\[
= [-4 \cdot [P_A' - v_A \cdot D'] - 4 \cdot [P_B' - v_B \cdot D'] + 2 \cdot [v_A + v_B] \cdot D']/\Omega_1
\] (5.37)

The marginal output effect is given by:

\[
\frac{dQ}{dt} = \frac{d\tau}{dt} \cdot \left( \frac{\partial q_A}{\partial w_A} + \frac{\partial q_A}{\partial w_B} + \frac{\partial q_B}{\partial w_A} + \frac{\partial q_B}{\partial w_B} \right) + \left( \frac{\partial q_A}{\partial w_B} + \frac{\partial q_B}{\partial w_B} \right)
\]

which can be simplified using Eq. (5.36) to:

\[
\frac{dQ}{dt} = 1 \cdot \left( \frac{\partial q_A}{\partial w_B} - \frac{\partial q_B}{\partial w_A} \right) = \frac{[v_A - v_B] \cdot D'}{2 \cdot \Omega_1}
\] (5.39)

where the last equality uses Eqs. (5.27)–(5.30).

The marginal effect on input prices follows from Eq. (5.37):

\[
\frac{dw_A}{dt} = \frac{d\tau}{dt} \cdot \frac{\partial q_A}{\partial w_A} = \frac{4 \cdot P_A' - [3 \cdot v_A - v_B] \cdot D'}{\Omega_2}
\]

\[
\frac{dw_B}{dt} = \frac{d\tau}{dt} + 1 = \frac{[3 \cdot v_B - v_A] \cdot D' - 4 \cdot P_B'}{\Omega_2}
\]

where $\Omega_2 = -4 \cdot [P_A' - v_A \cdot D'] - 4 \cdot [P_B' - v_B \cdot D'] + 2 \cdot [v_A + v_B] \cdot D' > 0$ (5.42)

From which Proposition 5.1 follows directly.

**Welfare analysis when price discrimination changes quantities in the same direction**

Consider that price discrimination changes both prices and both quantities in the same direction (either fall or rise). When the input price in both markets is higher than socially optimal under both pricing regimes, the bracketed terms of Eq. (5.15) are always positive and a sufficient condition for welfare improvement is that output increases in both markets. If the quantity decreases in both markets, price discrimination deteriorates welfare. Under price discrimination the input prices are higher than the welfare maximizing prices (see Eqs. (5.6) and (5.7)), but this is not necessarily true for the uniform price. Therefore the sufficient conditions for welfare improvement or deterioration in this case involve quantity changes (the sign of $dq_A/dt$ and $dq_B/dt$) but also conditions so that output is contracted in both markets under uniform pricing. The following proposition summarizes the sufficient conditions that characterize the welfare change in this case.

**Proposition 5.7.** When:

(a) Time valuations are sufficiently different in that $v_A - 5 \cdot v_B > -4 \cdot P_B'/D'$ and

139
(b) Congestion effects are not too high in that \( v_A \cdot D' < -5 \cdot P_B' + \sqrt{[5 \cdot P_B']^2 + 8 \cdot P_A' \cdot P_B'} \) or the time valuations are not too different in that \( v_B/v_A > [9 - \sqrt{73}]/4 \approx 0.114 \).

Price discrimination changes the quantity in both markets in the same direction and:

(i) Increases welfare if \( A_B/A_A > \lambda_1 \) (\( w_B > w_A \) holds and both quantities increase).

(ii) Decreases welfare if \( \lambda_0 < A_B/A_A < \lambda_1 \) (\( w_B < w_A \) holds and both quantities decrease), where \( \lambda_0 \) is defined in Appendix 5.B.

Proof: see below.

The conditions (a) and (b) imply that price discrimination changes both input prices and output in both markets in the same direction. That is, either both input prices fall and both quantities rise or vice versa. Proposition 5.7 (i) is intuitive. When the congestion effects and time valuations are such that conditions (a) and (b) in Proposition 5.7 hold, \( A_B/A_A > \lambda_1 \) implies that both prices fall and quantities rise. This, naturally, increases welfare because the input prices in both markets are higher than socially optimal in this case. Under these conditions, also the consumer surplus in each market increases. As discussed above, this can only occur when the ratio of reservation prices \( A_B/A_A \) is higher than the ratio of time valuations \( v_B/v_A \), which for conditions (a) and (b) to hold must be lower than 1/5. As a reference, \( \lambda_1 \) is greater than 1/2 when \( v_B/v_A = 1/5 \), so that the asymmetry of demand intercepts has to be significantly lower than the asymmetry of time valuations for welfare to increase with price discrimination.

In case (ii) of Proposition 5.7 the input prices rise and output falls in both markets, so that price discrimination necessarily decreases welfare if the prices were above socially optimal, which occurs when \( \lambda_0 < A_B/A_A \) (see Appendix 5.B for the definition of \( \lambda_0 \)). As in this case price discrimination increases the input prices and the differentiated prices are always higher than optimal (see Eqs. (5.6) and (5.7)), the uniform price is not necessarily higher than the socially optimal price of each market. For example, when \( q_B \) is relatively low and \( q_A \) is relatively high, the differentiated input price set by the facility in market B is similar to the socially optimal price. Therefore, as the uniform price is lower than the differentiated prices in this case, \( w \) can be lower than the socially optimal price for market B. This is likely to happen when \( A_B/A_A \) is sufficiently small, which explains why a lower bound for this ratio is needed for the sufficient condition (\( \lambda_0 < A_B/A_A \)). Again, as a reference, when \( v_B/v_A = 1/5 \), \( \lambda_0 < 1/2 \), so \( A_B/A_A = 1/2 \) is already small enough for the condition in Proposition 5.7 (ii) to hold.

Proof of Proposition 5.7

The first step of the proof is to show that \( \tau - q_A \cdot P_A' - q_B \cdot v_B \cdot D' \) and \( \tau + t - q_B \cdot P_B' - q_A \cdot v_A \cdot D' \) are always positive. From Eqs. (5.6) and (5.7) it follows that they are positive under price discrimination, so we need to show that they are positive for any value of \( t \) between zero and the optimal difference between input prices. As all functions are linear, the sign of the derivative of the terms does not change (see above that all derivatives are constant) and it is enough to show that the terms are positive under uniform pricing (at \( t = 0 \)). For \( \tau - q_A \cdot P_A' - q_B \cdot v_B \cdot D' \), we show that the term decreases with \( t \), which implies that
when \( A_B/A_A > \lambda_1 \) holds the price discriminatory price is approached by increasing \( t \) and therefore \( \tau - q_A \cdot P'_A - q_B \cdot v_B \cdot D' \) has to be positive at any \( t < t^* \):

\[
\frac{d}{dt} \left[ \tau - q_A P'_A - q_B v_B D' \right] = \frac{8P_A^2}{2} \left[ -P'_B + v_B D' \right] + [v_A - v_B] \left[ 12P_B v_A + [11v_A - v_B]v_B D' \right] D^2 + \left[ 2P'_A + 2P'_B - v_a - v_B \right] \left[ 4P_B^2 \left[ -P'_A + v_B \right] - v_B D' \right] - v_A v_B \left[ 4P'_A - v_B D' - v_B D' \right] D^2 < 0
\]

(5.43)

In the case where \( A_B/A_A < \lambda_1 \), the discriminatory price is approached by making \( t \) negative, so we need to assess directly \( w - q_A \cdot P'_A - q_B \cdot v_B \cdot D' \). Substituting the values of \( w \), \( q_A(w) \) and \( q_B(w) \), we obtain the following condition:

\[
w - q_A \cdot P'_A - q_B \cdot v_B \cdot D' > 0 \Leftrightarrow \frac{A_B}{A_A} \cdot F_B > -F_A \tag{5.44}
\]

where \( F_A = -2P'_B \left[ -3v_A + 11v_B \right] P'_A D' + v_B \left[ 7v_A + v_B \right] D^2 + 4P_B^2 - 4P_B^2 \left[ 3P_B - 2v_A D' \right] + v_B D' \left[ -5v_A + 12v_B \right] \left[ P'_A + v_B \left[ 7v_A + v_B \right] D^2 + 8P_B^2 \right] > 0 \)

\[F_B = P_B A \left[ -D' \left[ 2v_A D' + 2P_B \left[ 2v_A + 5v_B \right] + 2P_B^2 \left[ v_A D' + 2P_B \right] \right] \right] + 2P_B \left[ v_A D' + 2P_B \right] - v_A D' \left[ v_B v_B D' + 4 \left[ v_B - 2v_A \right] D' \right] \right]

If \( F_B > 0 \), the condition in Eq. (5.44) always holds. If \( F_B < 0 \), then the condition is equivalent to \( \frac{A_B}{A_A} < -F_A/F_B \), and as \( -F_A/F_B > \lambda_1 \) holds, \( A_B/A_A < \lambda_1 \) is sufficient.

Therefore, \( \tau - q_A \cdot P'_A - q_B \cdot v_B \cdot D' \) is positive for any value of \( t \).

For \( \tau + t - q_B \cdot P'_B - q_A \cdot v_B \cdot D' \) to be positive under uniform pricing, we assess its sign directly. Replacing \( w, q_A(w) \) and \( q_B(w) \), we obtain that it is positive when \( A_B/A_A > \lambda_0 \), where:

\[
\lambda_0 = \left[ P_B^2 \left[ 6v_A D^2 + 2[4v_A - 3v_B]P_B D' - v_B D^2 \right] + 2P_B^2 \left[ 2P_A + v_B D' \right] \right]
\]

\[
+ v_B \left[ 4v_B - 2v_A \right] P_A + v_B \left[ 5v_A - v_B \right] P_B \left[ 4v_B - 2v_A \right] D^2 \right]
\]

\[
\cdot \left[ 2P_B^2 \left[ 11v_A + 3v_B \right] P_B D' - v_B \left[ v_B + 7v_B \right] D^2 - 4P_B^2 \right] - 4P_B^2 \left[ 2v_B D' - 3P_B \right]
\]

\[
- v_A \left[ 11v_A + 3v_B \right] P_B D' - v_A \left[ v_A + 7v_B \right] D^2 + 8P_B^2 \right]^{-1}
\]

(5.45)

and, as \( \lambda_0 < \lambda_1 \), \( A_B/A_A > \lambda_1 \) is a sufficient condition for \( \tau + t - q_B \cdot P'_B - q_A \cdot v_B \cdot D' \) to be positive. In the case where \( A_B/A_A < \lambda_1 \), \( A_B/A_A > \lambda_0 \) is also needed.

The second part of the proof is to show that both quantities move in the same direction.
The marginal effect on downstream firm’s quantities are:

\[
\frac{dq_A}{dt} = \frac{\partial q_A}{\partial w_A} \cdot \frac{d\tau}{dt} + \frac{\partial q_A}{\partial w_B} \cdot \left[ \frac{d\tau}{dt} + 1 \right] \\
= \frac{2 \cdot \left[ P_B' - v_B \cdot D' \right] \cdot \left[ 4 \cdot P_A' - [5 \cdot v_A - v_B] \cdot D' \right] - v_A \cdot [v_A + v_B] \cdot D'^2}{\Omega_2} 
\]

\[
\frac{dq_B}{dt} = \frac{\partial q_B}{\partial w_A} \cdot \frac{d\tau}{dt} + \frac{\partial q_B}{\partial w_B} \cdot \left[ \frac{d\tau}{dt} + 1 \right] \\
= -\frac{2 \cdot \left[ P_A' - v_A \cdot D' \right] \cdot \left[ 4 \cdot P_B' - [5 \cdot v_B - v_A] \cdot D' \right] + v_B \cdot [v_A + v_B] \cdot D'^2}{\Omega_2} 
\]

From Eq. 5.47 it follows that part (a) of Proposition 5.7, \( v_A - 5 \cdot v_B > -4 \cdot P_B'/D' \), is sufficient for \( dq_B/dt > 0 \) as all other terms are positive.

As \( \Omega_2 > 0 \), we focus on the numerator of (5.46) to determine the sign of \( dq_A/dt \). Denote \( I \) the numerator and let \( v_B = \phi \cdot v_A \) where \( \phi \) is a constant in \([0, 1]\). This allows for focusing on a sufficient condition for \( v_B \) by taking into account that \( v_B < v_A \) must always hold. Solving \( I = 0 \) for \( v_A \), we get the following roots:

\[
r_1 = \frac{-P_B' \cdot [5 - \phi] - 4 \cdot P_A' \cdot \phi + \sqrt{[-P_B' \cdot [5 - \phi] - 4 \cdot P_A' \cdot \phi]^2 + 8 \cdot P_A' \cdot P_B' \cdot [1 - 9 \cdot \phi + 2 \cdot \phi^2]}}{D' \cdot [1 - 9 \cdot \phi + 2 \cdot \phi^2]} 
\]

\[
r_2 = \frac{-P_B' \cdot [5 - \phi] - 4 \cdot P_A' \cdot \phi - \sqrt{[-P_B' \cdot [5 - \phi] - 4 \cdot P_A' \cdot \phi]^2 + 8 \cdot P_A' \cdot P_B' \cdot [1 - 9 \cdot \phi + 2 \cdot \phi^2]}}{D' \cdot [1 - 9 \cdot \phi + 2 \cdot \phi^2]} 
\]

To prove that the condition in part (b) of Proposition 5.7 is sufficient for \( dq_A/dt > 0 \), we distinguish two cases. First, when \([1 - 9 \cdot \phi + 2 \cdot \phi^2] > 0 \), which is equivalent to \( \phi < [9 - \sqrt{73}] / 4 \approx 0, 114 \), \( r_2 \) is negative, \( r_1 \) is positive, and \( \partial^2 I / \partial v_A^2 < 0 \). Therefore for all values of \( v_A \) in \([0, r_1]\), \( dq_A/dt > 0 \). The minimum value of \( r_1 \) when \([1 - 9 \cdot \phi + 2 \cdot \phi^2] > 0 \) is achieved at \( \phi = 0 \), so that a sufficient condition is:

\[
v_A < r_1 \mid \phi = 0 = \frac{-P_B' \cdot 5 - \sqrt{[-P_B' \cdot 5]^2 + 8 \cdot P_A' \cdot P_B'}}{D'} 
\]

which is the condition in part (b) of Proposition 5.7. In the case where \([1 - 9 \cdot \phi + 2 \cdot \phi^2] < 0 \), which is equivalent to \( \phi > [9 - \sqrt{73}] / 4 \approx 0, 114 \), both roots are negative and \( \partial^2 I / \partial v_A^2 > 0 \) so that for all positive values of \( v_A \), \( dq_A/dt > 0 \) holds. This completes the proof that the condition \( v_A < r_1 \mid \phi = 0 \) or \( v_B/v_A > [9 - \sqrt{73}] / 4 \approx 0, 114 \) is sufficient for \( dq_A/dt > 0 \) to hold.

**Proof of Proposition 5.2**

- Welfare improvement: \( \frac{A_W}{A_A} > \lambda_1 \)

When \( v_A - 3 \cdot v_B < -4 \cdot P_B'/D' \) holds, it follows from Eqs. (5.40) and (5.41) that \( d\tau / dt < 0 \).
and \(dt/dt + 1 > 0\). This implies that \(dq_A/dt > 0\) and \(dq_B/dt < 0\) (see Eqs. (5.46) and (5.47)). From Eq. (5.15), it follows then that showing that \(w_A - q_A \cdot P_A - q_B \cdot v_B \cdot D'd\) which is positive (see Proof of Proposition 5.7), is greater than \(w_A + t - q_B \cdot P_B' - q_A \cdot v_A \cdot D'\) for any value of \(t \in [0, t^*]\) is sufficient for \(dW/dt > 0\). Denote \(f(t)\) the difference between these two terms; we prove that the condition in Proposition 5.2 is sufficient for \(f(t) > 0\) to hold. Formally,

\[
f(t) = -q_A \cdot \left[ P_A' - v_A \cdot D' \right] + q_B \cdot \left[ P_B' - v_B \cdot D' \right] - t
\]

\[
\frac{df}{dt} = -\frac{dq_A}{dt} \cdot \left[ P_A' - v_A \cdot D' \right] + \frac{dq_B}{dt} \cdot \left[ P_B' - v_B \cdot D' \right] - 1
\]

\[
\frac{df}{dt} = -\frac{8[P_A' + P_B'][-P_B' \cdot v_A \cdot D' + P_A' \cdot P_B' - P_B \cdot v_B \cdot D']}{\Omega_2}
\]

\[
\cdot [v_A - v_B]^2[v_A + v_B] \cdot D'^3 + [5v_A + v_B][-P_B' \cdot v_B - P_A' \cdot v_A]
\]

As \(df/dt < 0\), \(f(t^*) > 0\) is sufficient for \(dW/dt > 0\). Using that \(t^* = w_B - w_A\) and Eq. (5.34), we get:

\[
f(t^*) = \frac{A_A \cdot \left[ 12 \cdot P_A' \cdot P_B' + 10 \cdot v_A \cdot v_B + 2 \cdot v_B^2 + 3 \cdot v_B \cdot \left[ -4P_A' - P_B' \right] - 11 \cdot P_B' \cdot v_A \right]}{16 \cdot [-P_A' + v_A \cdot D'] \cdot [-P_B' + v_B \cdot D'] - [v_A \cdot D' + v_B \cdot D']^2}
\]

\[
- \frac{A_B \cdot \left[ 12 \cdot P_A' \cdot P_B' + 10 \cdot v_A \cdot v_B + 2 \cdot v_B^2 + 3 \cdot v_A \cdot \left[ -4P_B' - P_A' \right] - 11 \cdot P_A' \cdot v_B \right]}{16 \cdot [-P_A' + v_A \cdot D'] \cdot [-P_B' + v_B \cdot D'] - [v_A \cdot D' + v_B \cdot D']^2}
\]

from which the result in Proposition 5.2 follows directly, as the denominator is positive by the second-order conditions and the terms multiplying \(A_A\) and \(A_B\) in Eq. (5.55) are positive.

Finally, to show that the interval \([\lambda_1, \lambda_2]\) is non-empty when \(\frac{P_A'}{P_B'} \leq \frac{w_A}{v_B}\) holds, we look at \(\lambda_2 - \lambda_1:\)

\[
\lambda_2 - \lambda_1 = \frac{L_1 \cdot \left[ -P_B' v_A + P_A' v_B \right]}{L_2 \cdot L_3}
\]

where \(L_1 = 16 \cdot [-P_A' + v_A \cdot D'] \cdot [-P_B' + v_B \cdot D'] - [v_A \cdot D' + v_B \cdot D']^2 \geq 0\)

\(L_2 = 8 \cdot P_A' \cdot P_B' + 5 \cdot v_A \cdot v_B + v_B^2 + 2 \cdot v_A \cdot \left[ -4P_B' - P_A' \right] - 6 \cdot P_A' \cdot v_B \geq 0\)

\(L_3 = 12 \cdot P_A' \cdot P_B' + 10 \cdot v_A \cdot v_B + 2 \cdot v_B^2 + 3 \cdot v_A \cdot \left[ -4P_B' - P_A' \right] - 11 \cdot P_A' \cdot v_B \geq 0\)

where \(L_1 > 0\) follows from \(\partial^2 \Pi/\partial w_A^2 \cdot \partial^2 \Pi/\partial w_B^2 > \left[ \partial^2 \Pi/\partial w_A \partial w_B \right]^2\), a second-order condition that we assume to hold. Therefore, \(\lambda_2 - \lambda_1 \geq 0 \iff -P_B' v_A + P_A' v_B > 0\), which proves the result.

- Welfare deterioration: \(\frac{\Delta w}{\Delta w} < \lambda_1\)

As the price discriminating behavior is approached by making \(t\) negative in this case, the effect of price discrimination on welfare, output and prices have the opposite sign than
Chapter 5 Input third-degree price discrimination by congestible facilities

the marginal effect. That is, as \( w_A > w_B \) holds, welfare decreases when \( dW/dt > 0 \). As \( v_A - 3 \cdot v_B < -4 \cdot P'_B/D' \) holds, \( dq_A/dt > 0 \) and \( dq_B/dt < 0 \). Therefore, again, \( w_A - q_A \cdot P'_A - q_B \cdot v_B \cdot D' > w_A + t - q_B \cdot P'_B - q_A \cdot v_A \cdot D' \) for any value of \( t \in [-t^*, 0] \) is sufficient for \( dW/dt > 0 \) and thus for welfare deterioration. As \( df/dt < 0 \), the sufficient condition in this case is that \( f(0) > 0 \). Using Eqs. (5.24) and (5.25):

\[
f(0) = \frac{A_A \left[ -P'_B + v_B D' \right] \left[ -4P'_A + v_B D' + 5v_A D' \right] - A_B \left[ P'_A - v_A D' \right] \left[ 4P'_B - v_A D' - 5v_B D' \right]}{L_1}
\]

(5.57)

from which the condition \( \frac{A_B}{A_A} < \lambda_3 \) follows straightforwardly.

Appendix 5.C Calculations and proofs for Section 5.4

Input prices

Solving the first-order conditions for the public input supplier under price discrimination, \( \partial W_D/\partial w_A \) and \( \partial W_D/\partial w_B \), we get:

\[
w_A = \frac{A_B \cdot \left[ -2 \cdot P'_A \cdot (v_A + 2 \cdot v_B) \cdot D' + 4 \cdot v_A \cdot v_B \cdot D^2 \right]}{\left[ -P'_A + 2 \cdot v_A \cdot D' \right] \cdot \left[ -7 \cdot P'_B + 8 \cdot v_B \cdot D' \right] - 2 \cdot [v_A \cdot D' + v_B \cdot D']^2} - \frac{A_A \cdot \left[ 2 \cdot v_B \cdot (v_A + v_B) \cdot D^2 - P'_A \cdot [ -7 \cdot P'_B + 8 \cdot v_B \cdot D' \right]}{\left[ -P'_A + 2 \cdot v_A \cdot D' \right] \cdot \left[ -7 \cdot P'_B + 8 \cdot v_B \cdot D' \right] - 2 \cdot [v_A \cdot D' + v_B \cdot D']^2} - \frac{A_A \cdot \left[ -P'_B \cdot [4 \cdot v_A - 3 \cdot v_B] \cdot D' + 4 \cdot [v_A - v_B] \cdot v_B \cdot D^2 \right]}{\left[ -P'_A + 2 \cdot v_A \cdot D' \right] \cdot \left[ -7 \cdot P'_B + 8 \cdot v_B \cdot D' \right] - 2 \cdot [v_A \cdot D' + v_B \cdot D']^2} \]

(5.58)

\[
w_B = \frac{A_B \cdot \left[ -3 \cdot P'_B + 4 \cdot v_B \cdot D' \right] + v_A \cdot D'. \left[ -6 \cdot P'_B + 6 \cdot v_B - 2 \cdot v_A \cdot D' \right]}{\left[ -P'_A + 2 \cdot v_A \cdot D' \right] \cdot \left[ -7 \cdot P'_B + 8 \cdot v_B \cdot D' \right] - 2 \cdot [v_A \cdot D' + v_B \cdot D']^2} + \frac{A_B \cdot \left[ -P'_B \cdot [4 \cdot v_A - 3 \cdot v_B] \cdot D' + 4 \cdot [v_A - v_B] \cdot v_B \cdot D^2 \right]}{\left[ -P'_A + 2 \cdot v_A \cdot D' \right] \cdot \left[ -7 \cdot P'_B + 8 \cdot v_B \cdot D' \right] - 2 \cdot [v_A \cdot D' + v_B \cdot D']^2}
\]

(5.59)

Solving \( \partial W_D/\partial w \), we obtain:

\[
w = \left[ A_A \cdot \left[ 2[-P'_B + v_B D'][v_A D' - 4v_A D' - 6P'_A] + P'_A[-4P'_B + 7v_B D'] \right] - P'_B D^2[-2P'_B - P'_A] \right] + 2A_B \left[ -P'_B[-6P'_B + 8v_B D'][-P'_A + 2v_A D'] - P'_B[2v_A + v_B]D' + v_A D'[-P'_B[6v_A + v_B]D' + v_A D^2[7v_B - 2v_A] + 3P'_A v_A D'] \right] / \Omega_3
\]

144
where $\Omega_3 = v_B D' \left[\left[-4P_A' - v'_B D'\right] \left[\left[-8P_A' - P'_B + 16v_A D'\right] + 4v_A D' \left[7v_A D' + 3v_B D' - 3P'_B\right]\right] - 4P_A' P'_B \left[7P_A' + 2P'_B - 16v_A D'\right] - v_A D' \left[16P^2_B - 6P_A' v_A D' + 28P'_B v_A D' + 4v^2_A D'^2\right]\right]$

Effect of price discrimination on output and prices

To simplify notation, we again omit the arguments of the functions and let $\tau$ be the input price in market $A$ and $\tau + t$ the charge in market $B$. The marginal effect on $\tau$ is:

$$\frac{d\tau}{dt} = - \frac{\partial^2 W_D/\partial \tau \partial t}{\partial^2 W_D/\partial \tau^2}, \quad (5.60)$$

where,

$$\frac{\partial^2 W_D}{\partial \tau \partial t} = \left[\frac{\partial q_A}{\partial w_A} + \frac{\partial q_A}{\partial w_B}\right] \left[\frac{\partial q_A}{\partial w_A} + \frac{\partial q_A}{\partial w_B}\right] \left[P_A' - v_A D'\right] - v_A D' \left[\frac{\partial q_A}{\partial w_A} + \frac{\partial q_B}{\partial w_B}\right]$$

$$+ \left[\frac{\partial q_B}{\partial w_A} + \frac{\partial q_B}{\partial w_B}\right] \left[\frac{\partial q_B}{\partial w_A} + \frac{\partial q_B}{\partial w_B}\right] \left[\frac{1}{2} - \frac{\partial q_A}{\partial w_B} v_A D'\right] + \frac{\partial q_B}{\partial w_B}$$

Substituting Eqs. (5.27)–(5.30) in Eq. (5.60) yields:

$$\frac{d\tau}{dt} = \frac{4P^2_A' \left[-7P'_B + 8v_B D'\right] - P'_A \left[6v^2_A - 60v_A v_B + 8v^2_B\right] D'^2 - 6P'_B \left[v_B - 10v_A\right] D'}{\Omega_3}$$

$$+ 2v_A D' \left[-P'_B + v_B D'\right] \left[3v_B - 14v_A\right] D'^2 + v_A [2v_A + v_B] D'/\Omega_3$$

(5.63)

where $\Omega_3$ is positive as we assume that the second-order conditions of the supplier’s maximization problem under uniform pricing holds (i.e. $\partial^2 W_D/\partial w^2 < 0$). This, together with $v_A > v_B$ and $\partial^2 W_D/\partial w_A^2 \cdot \partial^2 W_D/\partial w_B^2 > [\partial^2 W_D/\partial w_A \partial w_B]^2$, which again holds as we assume that the second-order conditions of the supplier’s maximization problem under price discrimination holds, imply that $d\tau/dt < 0$. As $dw_A/dt = d\tau/dt$, we get that price discrimination decreases the input price in market $A$.

The marginal effect on $w_B$ is:

$$\frac{dw_B}{dt} = \frac{d\tau}{dt} + 1 = - \frac{4P'_B \left[2P'_A P'_B - v_A D' \left[4P'_B + P'_A\right]\right]}{\Omega_3}$$

$$+ \frac{v_B D' \left[-P'_A \left[-10P'_B + 4v_A D'\right] - P'_B \left[18v_A - v_B\right] D' + 2v_A [v_A + v_B] D'^2\right]}{\Omega_3}$$

(5.64)
which proves that price discrimination increases the input price in market $B$.

Using Eqs. (5.27)–(5.30), (5.38) and (5.63), we get:

$$
\frac{dQ}{dt} = \frac{2 \left[ P'_B - v_B D' \right] \left[ P'_A - [v_A - v_B]D' \right] + 2 \left[ P'_A - v_A D' \right] \left[ 4P'_B - [3v_B - v_A]D' \right] + P'_B v_B D'}{\Omega_3}.
$$

From Eq. (5.65) it follows that $v_A - 3 \cdot v_B < -4P'_B/D'$ is sufficient for $dQ/dt > 0$ to hold.

**Proof of Proposition 5.4**

Subtracting the value of the total welfare in Eq. (5.3) when evaluated at \{w_A, w_B\} and at $w$, we obtain:

$$
W(w_A, w_B) - W(w) = \frac{[w_B - w_A]}{2 \cdot \Omega_3} \cdot \frac{[A_A \cdot \lambda_4^N - A_B \lambda_4^D]}{\Omega_4},
$$

where

$$
\Omega_4 = 8P'_B [4P'_A + P'_B] [-P'_A + 2v_A D'] + 4P'_A P'_B + 2v_A^2 D'^2 [3P'_A - 14P'_B - 2v_A D']
$$

As $w_B > w_A$ holds, the sign of the welfare change is given by $[A_A \cdot \lambda_4^N - A_B \lambda_4^D]$, where $\lambda_4^N$ and $\lambda_4^D$ are given by Eqs. (5.67) and (5.68) respectively. As $\lambda_4^N$ is the numerator of $\lambda_4$ and $\lambda_4^D$ is the denominator of $\lambda_4$ and it is positive, the result proves Proposition 5.4.
\[
\lambda_4^N = \left[4v_A^6 + 7v_B^6 - 91v_A^2v_B^4 + 231v_A^2v_B^4 + 331v_A^3v_B^3 + 1350v_A^4v_B^2 - 198v_A^5v_B \right] D'^6 \\
+ \left[-3v_A^4 - 355v_Bv_A^3 + 6070v_B^2v_A^2 - 164v_B^3v_A + 116v_B^4 \right] P_A^2 D'^4 \\
+ 2v_B \left[35v_A^2 - 1568v_Bv_A + 32v_B^2 \right] P_A^3 D'^3 + 576v_B^2 D'^2 P_A^4 \\
+ \left[-4v_A^3 + 509v_Bv_A^4 - 4890v_B^2v_A^3 + 71v_B^3v_A^2 - 320v_B^4v_A + 62v_B^5 \right] P_A^4 D'^5 \right] P_B' \\
+ \left[432v_BD'P_A^3 + \left[-80v_A^5 + 5v_B^5 - 122v_A^4v_B + 439v_A^3v_B^2 + 554v_A^2v_B^3 + 1136v_A^3v_B \right] D'^5 \\
- \left[114v_A^3 - 4873v_B^2v_A^2 - 374v_B^3v_A - 216v_B^2 \right] P_A^2 D'^3 + \left[21v_A^2 - 2412v_Bv_A - 20v_B^2 \right] P_A^3 D'^2 \\
+ \left[192v_A^4 - 4092v_Bv_A^3 - 899v_B^2v_A^2 - 615v_B^3v_A + 76v_B^4 \right] P_A^4 D'^4 \right] P_B^2 \\
- \left[108P_A^3 - 160v_A^3 + 25v_B^3 - 186v_A^2v_B - 280v_A^3v_B \right] D'^5 \\
+ \left[488v_A^2 + 288v_Bv_A + 139v_B^2 \right] P_A' D'^2 - 4 \left[99v_A + 10v_B \right] P_A^2 D' \right] P_B^2 \left[P_A' - 2v_AD' \right] \\
- 4 \left[P_A' - 2v_AD' \right]^2 \left[3P_A' - 10v_AD' - 7v_BD' \right] P_B^3 \\
+ \left[6v_A^6D'^6 - 3 \left[3P_A' + 40v_BD' \right] v_A^5D'^5 + v_B \left[325P_A' + 538v_BD' \right] v_A^4D'^5 \right. \\
- v_B \left[1947v_BD'P_A' + 254P_A^2 + 124v_B^2D'^2 \right] v_A^3D'^4 \\
+ v_B \left[331v_B^2D'^2P_A' + 2490v_BD'P_A^2 + 56P_A^3 + 54v_B^3D'^2 \right] v_A^2D'^3 \\
- 4v_B^2 \left[4P_A' + v_BD' \right] \left[-4v_BD'P_A' + 84P_A^2 + 5v_B^2D'^2 \right] v_AD'^2 \\
+ 2v_B^2 \left[4P_A' + v_BD' \right]^2 \left[-2v_BD'P_A' + 8P_A^2 + v_B^2D'^2 \right] v_BD'^3 \right] (5.67)
\]
Chapter 5  Input third-degree price discrimination by congestible facilities

\[ \lambda^D = \left[ v_A \left[ 28v_A^5 - 80v_Bv_A^4 - 1075v_B^2v_A^3 - 247v_B^3v_A^2 + 155v_B^4v_A - 13v_B^5 \right] 
+ \left[ 110v_A^4 - 1187v_Bv_A^3 - 5083v_B^2v_A^2 + 160v_B^3v_A + 36v_B^4 \right] P_A^2D'^4 
- 5 \left[ 7v_A^4 - 174v_Bv_A^3 - 536v_B^2v_A + 16v_B^3 \right] P_A^3D'^5 - 16v_B \left[ 13v_A + 32v_B \right] P_A^4D'^6 
+ \left[ -100v_A^5 + 576v_Bv_A^4 + 4049v_B^2v_A^3 + 104v_B^3v_A^2 - 155v_B^4v_A + 6v_B^5 \right] P_A'D'^5 \right] P_B'D'^5 \]

\[ + \left[ -24v_A^4 - 5v_B^4 + 123v_Av_B^3 - 482v_A^2v_B - 468v_A^3v_B \right] D'^4 
- \left[ 223v_A^2 + 1248v_Bv_A + 152v_B^2 \right] P_A^2D'^2 + 4 \left[ 21v_A + 80v_B \right] P_A^3D' 
+ \left[ 144v_A^3 + 1478v_Bv_A^2 + 543v_B^2v_A - 72v_B^3 \right] P_A^3D'^3 \right] P_B' \left[ P_A' - 2v_AD' \right] 
- \left[ 68v_A^3D' - 4 \left[ 42P_A' - 53v_BD' \right] v_AD' + 5 \left[ 2P_A' - 5v_BD' \right] \left[ 6P_A' + v_BD' \right] \right] P_B' \left[ P_A' - 2v_AD' \right]^2 
-28 \left[ P_A' - 2v_AD' \right]^3 P_B'^4 \]

\[ + \left[ 2v_A^6 + 2v_B^6 - 8v_Av_B^5 - 94v_A^2v_B^4 + 412v_A^3v_B^3 + 34v_A^4v_B^2 - 28v_A^5v_B \right] D'^4 
+ \left[ -5v_A^3 - 18v_B^5 + 394v_Av_B^4 - 1573v_A^2v_B^3 + 269v_A^3v_B^2 + 103v_A^4v_B \right] P_A'D'^3 
+ \left[ 3v_A^4 - 412v_A^3v_B + 2088v_Av_B^3 + 597v_A^2v_B^2 - 120v_A^3v_B \right] P_A^2D'^2 
- \left[ 1200v_B^3 - 480v_Av_B^2 + 42v_A^2v_B \right] P_A^3D' + 128v_B^2P_A^4 \right] v_AD'^3 
\]

\[ + \left[ 16v_B^2D'^2P_A^2 + 128v_B^3D'P_A^3 + 256v_B^4P_A^4 \right] v_BD'^3 \] (5.68)

Appendix 5.D  Calculations and proofs for Section 5.5

The calculations and proofs are more brief in this section as they follow the same logic as the ones in the previous sections. Under downstream perfect price discrimination the derived demands are:

\[ q_A(w_A, w_B) = \frac{[-P_B' + 2 \cdot v_B \cdot D'] \cdot [A_A - w_A] - v_A \cdot D' \cdot [A_B - w_B]}{\Omega_1} \] (5.69)

\[ q_B(w_A, w_B) = \frac{[-P_A' + 2 \cdot v_A \cdot D'] \cdot [A_B - w_B] - v_B \cdot D' \cdot [A_A - w_A]}{\Omega_1} \] (5.70)

where \( \Omega_1 = [P_A' - 2 \cdot v_A \cdot D'] \cdot [P_B' - 2 \cdot v_B \cdot D'] - [v_A \cdot D'] \cdot [v_B \cdot D'] > 0 \) (5.71)
5. D Calculations and proofs for Section 5.5

Private facility

Solving the first-order conditions for the private input supplier under price discrimination we get:

\[
w_A = \frac{[-P'_A + 2 \cdot v_A \cdot D'] \left[ 2 \cdot A_A \cdot [-P'_B + 2 \cdot v_B D'] - A_B \cdot [v_A - v_B] D' \right] - A_A \cdot v_B [v_A + v_B] D'^2}{4 \cdot [-P'_A + 2 \cdot v_A \cdot D'] \cdot [-P'_B + 2 \cdot v_B \cdot D'] - [v_A \cdot D' + v_B \cdot D']^2}
\]

(5.72)

\[
w_B = \frac{[-P'_B + 2 \cdot v_B \cdot D'] \left[ 2 \cdot A_B \cdot [-P'_A + 2 \cdot v_A D'] + A_A \cdot [v_A - v_B] D' \right] - A_B \cdot v_A [v_A + v_B] D'^2}{4 \cdot [-P'_A + 2 \cdot v_A \cdot D'] \cdot [-P'_B + 2 \cdot v_B \cdot D'] - [v_A \cdot D' + v_B \cdot D']^2}
\]

(5.73)

\[
w = \frac{A_A \cdot [-P'_B + v_B \cdot D'] + A_B \cdot [-P'_A + v_A \cdot D']}{2 \cdot [-P'_A - P'_B] + 2 \cdot [v_A + v_B] \cdot D'}
\]

(5.74)

By subtracting both values we obtain that \(w_B - w_A > 0\) if and only if \(A_B/A_A > \lambda'_1\) where \(\lambda'_1\) is given by:

\[
\lambda'_1 = \frac{2 \cdot P'_A \cdot P'_B + 5 \cdot v_A \cdot v_B \cdot D'^2 + v_B^2 \cdot D'^2 + v_B \cdot D' \cdot \left[-4P'_A - P'_B\right] - 3 \cdot P'_B \cdot v_A \cdot D'}{2 \cdot P'_A \cdot P'_B + 5 \cdot v_A \cdot v_B \cdot D'^2 + v_B^2 \cdot D'^2 + v_A \cdot D' \cdot \left[-4P'_B - P'_A\right] - 3 \cdot P'_A \cdot v_B \cdot D'}
\]

(5.75)

Using the same methodology as in Appendix 5.B, we obtain:

\[
\frac{dQ}{dt} = \frac{[v_A - v_B] \cdot D'}{2 \cdot \Omega'_1} > 0 ,
\]

(5.76)

\[
\frac{dw_A}{dt} = \frac{2 \cdot P'_A \cdot [3 \cdot v_A - v_B] \cdot D'}{\Omega'_2} < 0 ,
\]

(5.77)

\[
\frac{dw_B}{dt} = \frac{[3 \cdot v_B - v_A] \cdot D' - 2 \cdot P'_B}{\Omega'_2},
\]

(5.78)

\[
\Omega'_2 = -2 \cdot \left[P'_A + P'_B - v_A \cdot D' - v_B \cdot D'\right] > 0
\]

(5.79)

which proves that when \(v_B/v_A \geq 1/3\), the prices move in opposite directions. As a consequence, quantities also move in the opposite direction with price discrimination.

The marginal welfare effect under downstream perfect price discrimination is:

\[
\frac{dW}{dt} = \frac{dq_A}{dt} \cdot [w_A - [q_B \cdot v_B \cdot D']] + \frac{dq_B}{dt} \cdot [w_A + t - [q_A \cdot v_A \cdot D']] ,
\]

(5.80)

Also in this case it is straightforward to show that the bracketed terms are positive at the discriminating input prices and that \(w_A - [q_B \cdot v_B \cdot D']\) decreases with \(t\), so that it is positive for all values of \(t\). Moreover, just as in Appendix 5.B the difference between the two terms, \(f'(t)\), decreases with \(t\). Then, when \(A_B/A_A > \lambda'_1\), \(f'(t^*) > 0\) is sufficient for
Chapter 5 Input third-degree price discrimination by congestible facilities

dW/dt > 0. Using the differentiated prices above, we obtain:

\[ f'(t^*) = \frac{A_A \cdot \left[-P_B' + 2v_B \cdot D' \right] \cdot \left[-2P_A' + [5v_A + v_B] \cdot D' \right]}{4 \cdot \left[-P_A' + 2 \cdot v_A \cdot D' \right] \cdot \left[-2P_B' + 2 \cdot v_B \cdot D' \right] - [v_A \cdot D' + v_B \cdot D']^2} \]

from which it follows that welfare increases when \( \lambda_1 < \frac{A_B}{A_A} < \lambda_2 \), where

\[ \lambda_2 = \frac{-P_B' + 2v_B \cdot D'}{\left[-2P_A' + [5v_A + v_B] \cdot D' \right]} \]

(5.82)

In the case where \( \lambda_1 > \frac{A_B}{A_A} \), \( f'(0) > 0 \) is sufficient for welfare to decrease with price discrimination by a private facility. Using the uniform price derived above we get:

\[ f'(0) = \frac{A_A \cdot L_A - A_B \cdot L_B}{\Omega_2 \cdot \left[P_A' - 2 \cdot v_A \cdot D' \right] \cdot \left[P_B' - 2 \cdot v_B \cdot D' \right] - [v_A \cdot D'] \cdot [v_B \cdot D']} \]

\[ L_A = \left[P_B' \left[2P_A' + P_B' - 3v_AD' \right] v_AD' + [5v_AD' \left[-P_B' + v_AD' \right] - P_A' \left[-P_B' + 4v_AD' \right] ] \right] v_BD' \]

\[ - \left[3P_A' + P_B' - 2v_AD' \right] v_B^2D'^2 + v_B^3D'^3 \]

\[ L_B = \left[\left[-P_A' \left[-P_B' + v_AD' \right] + v_AD' \left[-3P_B' + v_AD' \right] \right] v_AD' \right] + \left[\left[-P_A' + 2v_AD' \right] v_BD' \right] \]

\[ \left[\left[-P_A' - 2P_B' + 3v_AD' \right] v_BD' + \left[-3P_A' + 5v_AD' \right] v_B^2D'^2 \right] \]

from which it follows that welfare decreases when \( A_B/A_A < \min[\lambda_1, \lambda_3] \), where

\[ \lambda_3 = \frac{L_A}{L_B} \]

(5.84)

**Public facility**

Solving the first-order conditions for the public supplier, we obtain the following prices:

\[ w_A' = \frac{2 \cdot v_B \cdot \left[A_B \cdot \left[-P_A' + 2 \cdot v_A \cdot D' \right] - A_A \cdot \left[v_A + v_B \right] \cdot D' \right]}{\left[-P_A' + 2 \cdot v_A \cdot D' \right] \cdot \left[-3P_B' + 8 \cdot v_B \cdot D' \right] - 2[v_A + v_B]^2 \cdot D'^3} \]

(5.85)
\[ w' = \frac{A_A \cdot [-P'_B \cdot [2 \cdot v_A - v_B] \cdot D' + 4 \cdot [v_A - v_B] \cdot v_B \cdot D'^2]}{-P'_A + 2 \cdot v_A \cdot D'] \cdot [-3P'_B + 8 \cdot v_B \cdot D'] - 2[v_A + v_B]^2 \cdot D'^2} + \frac{A_B \cdot [-P'_A \cdot [-P'_B + 4 \cdot v_B \cdot D'] + 2 \cdot v_A \cdot D' \cdot [-P'_B + [3 \cdot v_B - v_A] \cdot D']]}{-P'_A + 2 \cdot v_A \cdot D'] \cdot [-3P'_B + 8 \cdot v_B \cdot D'] - 2[v_A + v_B]^2 \cdot D'^2} \quad (5.86) \]

\[ w = \left[ A_A \cdot [(P'_A + 2 \cdot v_A D')[-P'_B(2v_A - v_B) \cdot D' + 4v_B[v_A - v_B] \cdot D'^2] \right] \\
- v_B D'[-2P'_B v_A - P'_B v_B + 2v_A v_B \cdot D'] \cdot D' + A_B \cdot [[P'_A + 2v_A D'][P'_A - 2v_A D'] - v_A v_B D'^2 - 2v_A D'^2 - P'_B v_B D'] \\
- v_A v_B D'^2 P'_A] / \Omega_3 \quad (5.87) \]

where \( \Omega_3 = \left[ -P'_A + 2v_A D' \right] \left[ 3P'_A P'_B + 2P'^2 - 6P'_B v_A D' - 2v_A^2 D'^2 \right] + \left[ P'_B - 4v_A D' \right] v_B D'^2 + 2 \left[ -4P'^2 + 3P'_A P'_B - 14P'_A v_A D' - 4P'_B v_A D' + 14v_A^2 D'^2 \right] v_B D' \)

Subtracting the value of the total welfare under discriminating and uniform prices, we obtain:

\[ \Delta W = \frac{[w'_B - w'_A]}{\Omega_3} \cdot \left[ \frac{[A_A \cdot \lambda'_4^N - A_B \lambda'_4^D]}{4 \cdot \Omega_4^N} \right] \quad (5.88) \]

\[ \Omega_4^N = - \left[ 6P'_B v_A D' - P'_A \left[ 3P'_B - 8v_B D' \right] + 2 \left[ v_A^2 - 6v_A v_B + v_B^2 \right] \right] D'^2 \]

As \( w'_B > w'_A \) holds, it follows that the sign of the welfare change is given by \( [A_A \cdot \lambda'_4^N - A_B \lambda'_4^D] \) where \( \lambda'_4^N \) and \( \lambda'_4^D \) are defined below. As \( \lambda'_4^D \) is positive, the result proves that the sign of the welfare change has the same sign as \( A_B / A_A - \lambda'_4 \), where \( \lambda'_4 = \lambda'_4^N / \lambda'_4^D \).
\[
\lambda_A^{\prime \prime} = \left[ -2v_B P_A^{\prime 2} \left[ \frac{1}{2} \left( 24v_A^3 + 530v_B^3 + 132v_A^2v_B + 43v_B^2 \right) D^{\prime 4} + 2v_B P_A^{\prime 3} \left[ 49v_A + 13v_B \right] D^{\prime 3} \right.ight.
\]

\[
+ P_A^{\prime} \left[ 2v_A^5 - 183v_A^3v_B + 1946v_A^2v_B^2 + 309v_A^2v_B^3 + 262v_Av_B^4 - 52v_B^5 \right] D^{5} \\
- 2 \left[ 2v_B^6 + 89v_A^3v_B + 613v_A^2v_B^2 + 6v_A^2v_B^3 + 110v_Av_B^4 + 41v_A^5 + 11v_B^5 - 11v_B^6 \right] D^{6} \left]ight. \\
\left. + 2 \left[ 2P_A^{\prime 2} \left[ -8v_A^3 + 381v_A^2v_B + 146v_Av_B^2 + 62v_B^3 \right] D^{3} \\
+ P_A^{\prime} \left[ 64v_A^3 - 1448v_A^3v_B - 700v_A^2v_B^2 - 397v_Av_B^3 + 43v_B^4 \right] D^{4} - v_B P_A^{\prime 3} \left[ 133v_A + 25v_B \right] D^{2} \right]
\]

\[
+ 2 \left[ -32v_A^3 + 460v_A^2v_B + 226v_Av_B^2 + 158v_A^2v_B^2 - 39v_Av_B^3 + v_B^5 \right] D^{5} \left]ight. \\
\left. \right] \right] \right] \\
+ \left[ -2P_A^{\prime} \left[ 60v_A^3 + 72v_Av_B + 29v_B^2 \right] D^{2} + 3P_A^{\prime 2} \left[ 10v_A + v_B \right] D^{4} + 112v_A^3 D^{3} + 224v_A^2v_B D^{3} \\
+ 90v_Av_B^2 D^{3} - 9v_B^3 D^{3} \right] P_A^{\prime} \left[ P_A^{\prime} - 2v_A D \right] + 4 \left[ \left( 4v_A + v_B \right) D^{\prime} \left[ P_A^{\prime} - 2v_A D \right] \right] ^2 ] P_B^{\prime^4} \\
+ 8 \left[ 2v_B P_A^{\prime 2} \left[ -18v_A^3 + 242v_A^2v_B + 43v_Av_B^2 + 3v_B^3 \right] D^{4} - 32v_B P_A^{\prime 3} \left[ 3v_A + v_B \right] D^{3} \\
+ P_A^{\prime} \left[ -3v_A^5 + 129v_A^4v_B - 859v_A^3v_B^2 + 15v_A^2v_B^3 - 38v_Av_B^4 + 12v_B^5 \right] D^{5} \\
+ 2 \left[ 3v_A^5 - 60v_A^4v_B + 269v_A^3v_B^2 - 62v_A^2v_B^3 + 27v_Av_B^4 - 10v_A^5 + 15v_B^5 \right] D^{6} \left]ight. \\
\left. \right] \right] v_B D^{\prime} \quad (5.89)
\]
\[ \lambda_4^{(D)} = 4 \left[ 2v_B P_A^3 \left[ 12v_A^2 + 210v_A v_B + 43v_B^2 \right] D^3 - 52v_B^2 P_A^4 D^2 \right. \\
+ P_A^2 \left[ 7v_A^4 - 124v_A^3 v_B - 1269v_A^2 v_B^2 - 422v_A v_B^3 + 52v_B^4 \right] D^4 \\
+ 2P_A \left[ -14v_A^5 + 105v_A^4 v_B + 837v_A^3 v_B^2 + 324v_A^2 v_B^3 - 91v_A v_B^4 + 3v_B^5 \right] D^5 \\
+ 4v_A \left[ 7v_A^6 - 30v_A^5 v_B - 203v_A^4 v_B^2 - 77v_A^3 v_B^3 + 38v_A^2 v_B^4 - 3v_B^5 \right] D^6 \left. \right] P_B' \right. \\
- 2 \left. \left[ P_A^2 \left[ 11v_A^2 + 157v_A v_B + 124v_B^2 \right] D^2 - P_A \left[ 44v_A^3 + 354v_A^2 v_B + 454v_A v_B^2 - 43v_B^3 \right] D^3 \\
- 25v_B D' P_A^3 + [40v_A^4 + 256v_A^3 v_B + 393v_A^2 v_B^2 - 83v_A v_B^3 + 2v_B^4] D^4 \right] P_B' \left[ P_A' - 2v_A D' \right] \left. \right] P_B^3 \right. \\
- \left[ P_A' - 2v_A D' \right]^2 \left[ -2P_A [6v_A + 29v_B] D' + 3P_A^2 + [24v_A^2 + 116v_A v_B - 9v_B^2] D^2 \right] P_B^3 \left. \right] \left. \right] \left. \right] P_B' \left[ P_A' - 2v_A D' \right]^3 \right. \\
+ 8 \left. \left[ 32v_B^2 P_A^4 D^3 - 2v_B^2 P_A^3 \left[ 6v_A^2 + 124v_A v_B + 3v_B^2 \right] D^4 \\
- v_B P_A^2 \left[ 9v_A^4 - 48v_A^3 v_B - 705v_A^2 v_B^2 + 8v_A v_B^3 + 12v_B^4 \right] D^5 \\
+ P_A' \left[ -v_A^5 - 31v_A^4 v_B + 67v_A^3 v_B^2 + 881v_A^2 v_B^3 - 86v_A v_B^4 - 26v_A v_B^3 + 2v_B^5 \right] D^6 \right) \right. \\
+ 2v_A \left[ v_A^6 - 14v_A^5 v_B + 17v_A^4 v_B^2 + 206v_A^3 v_B^3 - 47v_A^2 v_B^4 - 4v_A v_B^5 + v_B^6 \right] D^7 \right] (5.90) \]
Chapter 6

Airline route structure competition and network policy
6.1 Introduction

Following the deregulation of the airline industry, several changes in aviation markets were observed (see Morrison and Winston (1995) for an overview of the changes in the US industry, and Burghouwt and Hakfoort (2001) for Europe). In addition to changes in fares, the most notorious change was in the way markets were served: the adoption of hub-and-spoke route structures by carriers became dominant. Also since the deregulation of the markets, the costs caused by congestion at airports have grown significantly and managing the increasing congestion has become one the main concerns of governments in countries with a large aviation market. For example, Ball et al. (2010) estimate that the cost of US air transportation delay in 2007 was $16.7 billions to passengers, $8.3 billions to air carriers and that it reduced the 2007 GDP by $4 billion. Not surprisingly, optimal airport pricing has gained increased attention as a measure aimed at reducing congestion costs.

The objective of this chapter is to analyze optimal airport pricing in a network setting and in the presence of congestion externalities, where carriers with market power have the route structure choice as a strategic instrument. It is known from earlier literature, which abstracts away from endogenous route structure, that oligopolistic carriers partially internalize congestion and exert market power (e.g., Brueckner, 2002; Pels and Verhoef, 2004). This means that two inefficiencies need to be corrected: the deadweight loss from market-power markups (e.g., with subsidies) and the excessive number of flights that are scheduled (e.g., with slot constraints or congestion pricing). In this chapter, we study whether and how the inclusion of route structure choice by carriers changes these conclusions. Specifically, do regulators need an additional instrument, on top of the ones described above, to induce the socially desirable outcome? We carry out the analysis in what we believe is the simplest possible setting that allows us to account for strategic interactions in route structure choice, endogenous hub locations, market power exertion by airlines, congestion externalities at airports, and passenger frequency benefits and transfer costs.

The conditions that give rise to hub-and-spoke route structures as an equilibrium of unregulated competition have often been explained with three arguments: economies of density, frequency effects and strategic advantages. The first refers to that average cost in a direct route may decrease with the number of passengers, and the second, the frequency effect, to the fact that there are benefits for passengers of increased frequencies, e.g., reductions in schedule delay costs (the difference between desired and actual departure/arrival time). Both may be better exploited under hub-and-spoke structures. In a monopoly framework, Hendricks et al. (1995) show that economies of density alone can induce an airline to adopt a hub-and-spike route structure; Brueckner (2004) shows how frequency effects favor the adoption of hub-and-spike. The third argument, strategic advantages, reflects that adopting hub-and-spoke route structures may bring, in oligopolistic competition, further advantages because of the effect it has on competitors. For instance,
Oum et al. (1995) show that using a hub-and-spoke structure may allow the carrier to be more aggressive in output market competition. Employing hub-and-spoke can deter entry in hub markets if the complementarities among hub markets are large, or the number of complementary hub markets is large (Hendricks et al., 1997). Finally, it may prevent competition in local markets between two hub carriers because invading the competitor’s local market may reduce own profit in all connecting markets due to a more aggressive competition in the trans-hub market (Zhang, 1996). Recent theoretical contributions to this topic include Hendricks et al. (1999), Alderighi et al. (2005), Barla and Constantatos (2005), and Flores-Fillol (2009, 2010).63

The above studies, however, ignore the endogenous nature of the hub location, do not study the socially optimal route structure, and most of them also ignore congestion effects. This chapter contributes to the literature by including all these together. For example, we find that airlines choosing hub-and-spoke structures using different cities as their hub may be the unique equilibrium when airports, markets and airlines are symmetric. The consideration of asymmetric hub-and-spoke networks, commonly observed in real markets, is an important contribution of this chapter. In addition, we show that the result that a monopolistic airline is biased towards hub-and-spoke configurations (Brueckner, 2004) does not necessarily carry over to competing airlines under our assumptions. We find that airlines adopting fully connected route structures can be the unique equilibrium when using hub-and-spoke structures is socially optimal.

The literature on pricing and regulation in aviation markets has mostly focused on either a single origin destination pair, hence ignoring network effects, or in networks with fixed route structures, hence ignoring its endogenous nature and its effect on optimal policy. Brueckner (2004) compares the optimal route structure in a three-node network with the one adopted by an unregulated monopoly, but does not study pricing policies. Flores-Fillol (2009) extends the analysis to a duopoly of airlines without analyzing the optimal route structure. Brueckner (2005) and Flores-Fillol (2010) study the optimal pricing policy in a duopoly setting with an exogenous route structure. Our contribution to the policy analysis is to identify the rationale for the tolls, when route structure is endogenous, and to extend the optimal pricing and regulation analysis by elaborating on the policy instruments that can decentralize the socially efficient outcome in terms of output and network configuration.

A main result of our analysis is that a regulatory instrument directly targeted on route structure choice may be needed to maximize welfare, in addition to tolls designed to induce the efficient outputs, given the networks chosen. We find that social welfare can be increased by using an additional policy instrument when the regulator is restrained from subsidizing airlines (needed to eliminate deadweight losses), but also when it does not face such constraint on tolling. Specifically, the first-best optimal route structures and output levels cannot always be decentralized by just using an airline- and market-specific per-passenger toll (to correct for market power), together with an airline- and link-specific per-flight toll (to correct for congestion), designed to induce the efficient output for the optimal route structure. Thus, the equilibrium with those tolls is not always efficient, even when the regulator can perfectly discriminate airlines and has no pricing constraints. This

63For an empirical analysis and review of the size and shape of the networks in the airline industry from a cost perspective, see Jara-Díaz et al. (2013).
is because these tolls, despite that they provide the incentives to set the output efficiently, cannot always align the effect of adopting different route structures on the firms’ profit with the effect on social welfare of those different configurations. This is especially true when the optimal network configuration is asymmetric and requires one firm to have, in one of the markets, a significantly lower market share and profit than the competitor.

First-best pricing, as just discussed, typically requires a regulator to give per-passenger subsidies to airlines, a policy that is arguably impossible to implement in practice. To address this limitation, we study the case in which the regulator is constrained to charge non-negative tolls. We show that the route structures and output levels that maximize welfare in absence of subsidies to correct for market power exertion cannot always be decentralized through non-negative tolls alone. Thus, also in the absence of subsidization, using an additional regulatory instrument, on top of the tolls designed to correct output inefficiencies, may increase welfare.

Our results may have important policy implications. In some cases an instrument directly aimed at regulating route structure choice is needed for welfare maximization, and in the cases where the pricing instruments are sufficient, the rationale for the charges is not always the same. In some cases they are required only to correct output choices, in other cases the tolls are needed to correct simultaneously output and route structure choices, and finally they can also be needed in order to change the market structure in terms of suppliers present in the network, in addition to correct output and route structure. It is therefore evident from the analysis that a regulator designing a pricing policy for the aviation industry has to take into account its effect on the long-run route structure equilibrium, and assess the optimality of the observed setting before deciding on toll levels. This is particularly relevant for regions in which all airports are owned and operated by the same authority, and puts a question mark on the use of airport price-fixing systems, where all airports in a network charge identical tolls to airlines. As Bel and Fageda (2010) point out, in Spain, Greece (except Athens) and Norway there is a single operator that applies this system approach. Moreover, airlines operating in the local (national) markets in those countries operate asymmetric hub-and-spoke structures with different airports as their hubs, something that highlights our methodological contribution of allowing for these cases.

In addition to the policy implications discussed above, we show that more information is needed to achieve the regulator’s objective: under fixed network structures, adaptive pricing based on “local variables” (notably marginal external cost and marginal benefit) would “guide” the regulator to the optimum. But, with discrete changes in route structures, and multiple local optima, this is no longer true. A “naive” regulator, who observes a sub-optimal route structure configuration, and sets the tolls based on this, may not always achieve welfare maximization, neither first-best nor second-best.

The remainder of the chapter is structured as follows. In the next section we introduce the model and the configuration of nodes between which endogenous route structures can be offered. Section 6.3 analyzes the equilibria of the untolled competition for a duopoly, while Section 6.4 derives the welfare-maximizing combination of output and route structure. Section 6.5 studies whether pricing is enough to achieve the welfare-maximizing setting as the result of tolled competition, and extends the analysis to the second-best case where the regulator cannot subsidize airlines. Finally, Section 6.6 concludes.
6.2 The model

In order to keep the simplest possible focus on the route structure choice by agents with market power in presence of externalities, we use a stylized model that follows Brueckner’s (2004) in the basic assumptions, and extends it by considering congestion, airline competition and the analysis of how to enforce the social optimum.

We consider a symmetric duopoly of airlines that compete in each of the three symmetric markets that are shown in Figure 6.1. These markets $M = \{XY, YZ, XZ\}$ represent return-trips for simplicity (e.g. people travel from X to Y and return). The links $L = \{xy, yz, xz\}$ are always available to any airline; that is, both airlines have permission to schedule flights in any city-pair. Each market $m$ can be served by airlines either directly, flying nonstop from the origin airport to the destination airport, or via a hub airport that an airline chooses to use for the connection. As a result, the two possible route structures for an airline are: fully connected (henceforth F); or hub-and-spoke (henceforth H), where they choose one airport as its hub, and fly only between the hub and the two remaining airports, serving two markets non-stop and one with connecting flights. As we let each airline’s hub be endogenous, asymmetric settings with hub-and-spoke structures may arise.

![Figure 6.1: Network](image)

We model the airlines’ competition with a two-stage game where, first, carriers simultaneously choose route structure, and then they compete in output at a market level. Specifically, in the latter stage, airlines have the number of passengers in each market ($q_m$), the number of flights in each link ($f_l$), and the aircraft size ($s_l$) as strategic variables. This is an extension of the Cournot assumption that airlines take rival’s quantities, instead of fares, as given.

We assume that the full price faced by a passenger traveling nonstop with airline $i$, is:

$$\theta_i^m = p_i^m + D_i^m + g_i^m.$$  \hspace{1cm} (6.1)

---

64 Direction-specific return-trips (e.g. people making round-trips both from X to Y and from Y to X) may be introduced without altering the analysis.
65 Equilibria where an airline chooses to serve fewer than the 3 origin-destination pairs is considered in the numerical model.
66 The assumption that airlines compete in a Cournot fashion is common in the airline literature and is supported by empirical evidence by Brander and Zhang (1990) and Oum et al. (1993). For a discussion of the implications of leadership behavior see Brueckner and Van Dender (2008), and for Bertrand competition with differentiated airlines see Silva and Verhoef (2013).
This is the sum of the fare, \( p^m_i \), the congestion delay cost, \( D^m_i \), and schedule delay cost, \( g^m_i \). For a connecting trip, passengers incur an additional cost of layover time that we assume fixed and equal to \( \mu \). We further assume that airlines are perceived as imperfect substitutes and that the passenger demand function for an airline \( i \) in market \( m \), \( q^m_i \), is linear in the own and in the rival’s price. Therefore, the demand faced by the airline depends on its own full price, \( \theta^m_i \), as well as its rival’s, \( \theta^m_j \) (hereafter, when subscript \( j \) appears in the same expression with \( i \), it refers to the rival airline). These assumptions are summarized in the following inverse demand:

\[
\theta^m_i = A - B \cdot q^m_i - E \cdot q^m_j ,
\]

where \( A, B, \) and \( E \) are positive parameters satisfying \( 0 \leq E \leq B \). Note that we ignore demand dependencies between markets (city-pairs). This set of assumptions allows us to analyze the effect of airline horizontal differentiation on route structure choice by means of varying the ratio of substitutability, \( E/B \), that ranges from 0, when the airlines’ outputs are independent in terms of demand interaction, to 1, when the airlines’ outputs are perfect substitutes. At the same time, we consider that not all passengers choose the airline with the most attractive fare-delay combination due to factors that may differ across carriers, such as service level (e.g. language), and make passengers perceive airlines as imperfect substitutes.

Following Brueckner (2004), we model the airlines’ cost per flight as a function of the aircraft size \( (s) \) as:

\[
C(s) = c_f + c_s \cdot s ,
\]

where \( c_f \) is the fixed cost per flight, and \( c_s \) the marginal cost per seat. This formulation captures in a simple way that increasing the number of passengers per link may reduce average cost per passenger through economies of seats. We also assume a constant load factor of 100\%, which allows for analytical tractability. A more realistic model would have endogenous aircraft size and load factors, with stochastic demand, but this would prevent us from providing transparent understanding of route structure equilibria. We assume that congestion affects airlines only through the demand function, in that increased congestion raises the full price faced by passengers and therefore, everything else constant, fares will be lowered by the increased congestion. We could also include that congestion affects the airlines’ costs, but this would not affect the main results and conclusions in any fundamental way. In the profit function, there is no difference if congestion raises the airlines’ costs or if reduces the passengers’ willingness to pay.

As a natural benchmark, we consider a regulator that controls all airports and maximizes welfare, so that we analyze a three-stage game. In the first stage, the regulator sets per-passenger tolls to each airline in each market \( (\tau^m_i) \), and per-flight tolls to each airline in each link \( (\tau^l_i) \). In the second and third stage, airlines choose route structure and

---

67 The schedule delay is the time difference between a passengers desired departure time and the actual departure time. As we do not explicitly model trip-timing decisions, this represents an average measure of the schedule delay cost, that, at least, provides the right intuition.

68 This formulation is a reduced form of a system of welfare-maximizing airports, or a regulator, setting charges per passenger and charges per flight at each airport. The difference would matter if each airport is controlled by a different authority, that may, for instance, maximize local welfare.
output respectively.\(^{69}\) We look at subgame-perfect Nash equilibria through backward induction, so we first analyze, in the following section, the airlines’ output Nash equilibrium for a given network configuration and then we study the route choice Nash equilibrium (where airlines consider the effects of their network choice on the own and competitor’s outputs).

### 6.3 Airlines equilibrium

In the second stage of the game, airlines have a discrete choice between alternative route structures (\(F\) or \(H\) at any of the airports). For this reason, we first look at their profits while taking route structures as given, and then analyze the equilibrium in route structure.

#### 6.3.1 The fully connected route structure

In this setting, airline \(i\) uses \(F\) as its route structure, and chooses the frequency on each link, \(f^i_l\), and the number of passengers in each market, \(q^m_i\). Due to our assumption that the load factor is constant and equal to 100\%, the seats per flight are \(s^i_l = q^i_l / f^i_l\), where \(q^i_l\) is the number of passengers flying through link \(l\).

We assume that the average schedule delay depends only on the flight frequency of the airline in the link that connects that market, and that it decreases with frequency (e.g. \(\partial g^i_{XY} / \partial f^i_{XY} < 0 \land \partial g^i_{XY} / \partial f^i_l = 0 \forall l \neq XY \land \partial g^i_{XY} / \partial f^i_j = 0 \forall l \neq ij\)). The assumption that schedule delay does not depend on the rival’s frequency, as congestion does, reflects our view that, in the differentiated duopoly, frequency is perceived as an airline-specific attribute. We also assume that there is congestion at the origin and at the destination, that the airport runway congestion depends on the total number of flights at that airport, and that it increases in the total number of flights. For example, denoting \(F^l = f^i_l + f^j_l\) the total number of flights on the link \(l\), the full price faced by a passenger of market \(XY\) flying with airline \(i\) is:

\[
\theta^i_{XY} = p^i_{XY} + D(F^x + F^z) + D(F^x + F^z) + g^i_{XY}(f^i_{xy}) ,
\]

where \(D\) is the delay cost function, assumed common to all airports. The full price equals the sum of the following: the fare; the congestion at the origin, \(X\), which depends on the total number of flights operating at that airport \(F^x + F^z\); the congestion at the destination airport, \(Y\); and the schedule delay cost. Without loss of generality, both delay functions, \(D\) and \(g\), include the passengers valuation of time and reflect that these costs are experienced also in the return trip. We look at the case where all airports have the same delay function, but this could easily be extended.

The airline’s profit, using Eq. (6.3) and \(s^i_l = q^i_l / f^i_l\), is:

\[
\pi^F_i = \sum_{m \in M} q^m_i \cdot (p^m_i - c_q - \tau^m_i) - \sum_{l \in L} f^i_l \cdot (c_f + \tau^i_l) ,
\]

\(^{69}\)We assume that airlines treat the tolls as parametric, i.e. they do not believe that their actions may change the way a regulator chooses instruments. See Brueckner and Verhoef (2010) for a detailed discussion on how to account for such behavior of agents with market power in presence of externalities.
where the superscript $F$ refers to the fully connected structure. Using this, together with 
Eq. (6.2), we can rewrite profit as:

$$
\pi^F_i = \sum_{m \in M} q^m_i \cdot (A - B \cdot q^m_i - E \cdot q^m_i - D^m_i - c_q - \tau^m_i) - \sum_{l \in L} f^l_i \cdot (c_f + \tau^l_i) .
$$

(6.6)

The airline’s profit indirectly depends also on the rival’s route structure. That structure will be reflected in the rival’s number of passengers and flights, which will affect demands and delays. What we do, in this section, is to look at the airlines’ best response in output irrespective of which route structure underlies the rival’s quantities, and then, when deriving the equilibrium in route structure, compare the differences that arise from the different rival’s route structure choices. The first-order conditions for $q^m_i$ and $f^l_i$ imply the following pricing and frequency setting rules:

$$
\frac{\partial \pi^F_i}{\partial q^m_i} = 0 \Rightarrow p^m_i = c_q + \tau^m_i + B \cdot q^m_i ,
$$

(6.7)

$$
\frac{\partial \pi^F_i}{\partial f^l_i} = 0 \Rightarrow -\sum_{m \in M} q^m_i \cdot \left( \frac{\partial D^m_i}{\partial f^l_i} + \frac{\partial g^m_i}{\partial f^l_i} \right) = c_f + \tau^l_i .
$$

(6.8)

Eq. (6.7) states that the fare charged by the airline in market $m$ is the sum of the marginal cost per capacity unit ($c_q$), the airport charge per passenger in that market ($\tau^m_i$), and a conventional markup reflecting the carrier’s market power ($B \cdot q^m_i$). Eq. (6.8) states that the airline’s marginal cost per flight (right-hand side of Eq. (6.8)) equals marginal revenue (left-hand side, marginal congestion costs plus marginal schedule delay benefits). Therefore, airlines internalize own-passenger congestion. These rules are analogous to the rules obtained previously in Cournot competition (e.g. Pels and Verhoef, 2004).

The airline’s profit using a fully connected route structure, in sub-game equilibrium, is:

$$
\Pi^F_i = \sum_{m \in M} B \cdot (q^m_i)^2 - \sum_{l \in L} f^l_i \cdot (c_f + \tau^l_i) ,
$$

(6.9)

which is obtained by replacing Eq. (6.7) into Eq. (6.6). This is the revenues from the markup ($B \cdot q^m_i$ per passenger) minus the constant per-flight costs.

### 6.3.2 The hub and spoke route structure

We now look at the case where airline $i$ chooses to serve the markets with a hub-and-spoke route structure. For illustration purpose, we analyze the case where the airline chooses airport $Y$ as its hub. Other cases are simply obtained by changing notation only. When we study the full game equilibrium, then it is necessary to explicitly model the choice of hub airport. The change with respect to the fully connected case is that the market $XZ$ (the spoke market) is served with connecting flights at the hub. We assume, as in previous studies of hub-and-spoke route structures, that the fare in the spoke market is set independently; this implies that the fare in market $XZ$ is not restricted to be equal to the sum of the fares of the two hub markets ($XY$ and $YZ$). The fares must, however,
6.3 Airlines equilibrium

satisfy an arbitrage condition: the sum of the fares of the hub markets (in this case, $XY$ and $YZ$) cannot be lower than the fare charged to the connecting passengers (market $XZ$).\(^{70}\)

The number of seats per flight changes in this case because, in addition to the passenger from hub markets, the passengers from the spoke markets are also traveling through links $xy$ and $yz$. As a consequence, in this setting, aircraft sizes will satisfy:

\[
\begin{align*}
    s_{xy}^i &= \frac{(q_{XY}^i + q_{XZ}^i)}{f_{xy}^i}, \\
    s_{yz}^i &= \frac{(q_{YZ}^i + q_{XZ}^i)}{f_{yz}^i}.
\end{align*}
\]

(6.10)

Full prices in the hub markets have the same structure as before (see Eq. (6.4)), but they change in the spoke market. We assume that the passengers’ congestion delay, denoted $\tilde{D}_{XZ}^i$, is the sum of the delays at each leg and that those passengers incur a constant transfer cost of $\mu$, a standard assumption in the literature (e.g. Brueckner, 2004). The schedule delay of a connecting passenger is not straightforward to represent, in average terms, as a function of frequencies. Brueckner (2004) assumes that it is equal to the schedule delay of a passenger flying only one leg. In our model there is imperfect competition and endogenous hub location, therefore, that assumption would only be applicable if we impose \textit{ex ante} symmetry. To avoid this, and for practical purpose, we suppose that the schedule delay cost for a connecting passenger, denoted $\tilde{g}_{XZ}^i$, can be modeled as the average of the schedule delays in each of the two direct routes.\(^{71}\) These assumptions give:

\[
\begin{align*}
    \tilde{D}_{XZ}^i &= D_{XY}^i + D_{YZ}^i, \\
    \tilde{g}_{XZ}^i &= (g_{XY}^i + g_{YZ}^i)/2, \\
    \theta_{XZ}^i &= p_{XZ}^i + \tilde{D}_{XZ}^i + \tilde{g}_{XZ}^i + \mu.
\end{align*}
\]

(6.11)

The above assumptions shape profit in the following way:

\[
\begin{align*}
    \pi_i^H &= \sum_{m \in \{XY,YZ\}} q_m^i \cdot (A - B \cdot q_m^i - E \cdot q_j^m - g_{i}^m - D_{i}^m - c_q - \tau_m^i) + \\
    &+ q_{i}^{XZ} \cdot (A - B \cdot q_{i}^{XZ} - E \cdot q_j^{iXZ} - \tilde{g}_{i}^{XZ} - \tilde{D}_{i}^{XZ} - \mu - 2 \cdot c_q - \tau_{i}^{XZ}) - \sum_{l \in \{xy,yz\}} f_l^i \cdot (c_f + \tau_l^i),
\end{align*}
\]

(6.12)

where the superscript $H$ refers to an airline serving the markets with a hub-and-spoke route structure. The difference, besides the new definition of delays, is that, everything else constant, an additional passenger in the $XZ$ market requires an increase of aircraft

\(^{70}\)Note that we do not impose the arbitrage condition that the fare charged to connecting passengers has to be higher than the fare charged to passengers in direct markets. Instead, we assume that airlines can effectively prevent a passenger that purchases a ticket in the connecting market to make a return trip from the hub airport to one of the other two airports. As we model return trips, it is a reasonable assumption because, at least in one of the two directions, the passenger that intends to deviate will have to board a flight at the hub with a ticket that reveals the purchase of a trip originating in a different city. For the case of non-return trips, this is less likely and the additional arbitrage condition would have to be imposed.

\(^{71}\)An alternative is to use the minimum frequency considering the frequency on the two routes, but this would generate unnecessary analytical complications to the model without changing results in any significant way.
size in both links, making the cost per connecting passenger equal to \(2 \cdot c_q\).

The first-order conditions lead to the following pricing and frequency setting rules:

\[
\frac{\partial \pi_i^H}{\partial q_{im}^m} = 0 \Rightarrow p_{im}^m = c_q + \tau_{im}^m + B \cdot q_{im}^m \quad \forall m \in \{XY, YZ\},
\]

(6.13)

\[
\frac{\partial \pi_i^{XZ}}{\partial q_{im}^{XZ}} = 0 \Rightarrow p_{i}^{XZ} = 2 \cdot c_q + \tau_{i}^{XZ} + B \cdot q_{i}^{XZ},
\]

(6.14)

\[
\frac{\partial \pi_i^H}{\partial f_{il}^l} = 0 \Rightarrow -\sum_{m \in \{XY,YZ\}} (q_{im}^m + q_{i}^{XZ}) \cdot \left( \frac{\partial D_{l}^m}{\partial f_{il}^l} + \frac{\partial g_{m}^i}{\partial f_{il}^l} \right) = c_f + \tau_{i}^l \quad \forall l \in \{xy,yz\}.
\]

(6.15)

These equations state that airlines apply a market power markup in each market and set frequency to equalize marginal revenue with marginal cost, hence partially internalizing congestion. In this setting, the sub-game equilibrium profit, using the first-order conditions (Eqs. (6.13) and (6.14) in Eq. (6.12)), can be written as:

\[
\Pi_i^H = \sum_{m \in M} B \cdot (q_{im}^m)^2 - \sum_{l \in \{xy,yz\}} f_{il} \cdot (c_f + \tau_{i}^l).
\]

(6.16)

Just as in the previous case, it is the revenues from the markup minus frequency costs.

### 6.3.3 Second-stage: the choice of route structure

In contrast to the third-stage, in this stage the airlines’ decision variables are discrete. Airlines either serve the markets with a fully connected route structure or with a hub-and-spoke route structure. As we ignore the possibility of serving only one market in the analytical model, no other configurations are possible. When an airline chooses to use a hub-and-spoke structure, it also chooses which airport to use as the hub. To characterize the equilibria, we need to compare the airlines’ best responses, knowing the outcome of the third stage (quantity and frequency), for all the rival’s possible route structures.

Denote the route structure choice of an airline \(i\) as a choice of \(r_i \in RS = \{F, H_X, H_Y, H_Z\}\), where \(H_C\) refers to a hub-and-spoke structure with airport \(C\) as the hub. The relevant comparisons are between the profits under different route structures, given the route structure of the rival. Let \(\Pi_i(r, v)\) be the profit of airline \(i\), evaluated at the outputs of the third-stage equilibrium, when it has chosen \(r\) as its route structure, and the rival uses \(v\). Then, it follows that a symmetric setting with both airlines using route structure \(r\) will be an equilibrium of the airlines’ game if and only if:

\[
\Pi_i(r, r) \geq \Pi_i(u, r) \quad \forall u \in RS.
\]

(6.17)

Because airlines are symmetric, whenever this holds true for one airline, it will hold true for the other as well, and both airlines having \(r\) will be a subgame equilibrium of the airlines’ competition. Also note that, whenever both airlines playing \(H_X\) is an equilibrium, both having \(H_Y\) and both having \(H_Z\) are also equilibria, because airports and markets are symmetric as well. We will refer to this set of equilibria as \((H, H)\): both airlines using hub-and-spoke route structures and both using the same airport as their hub. It
follows that \((F, F)\) is the equilibrium where both airlines choose the fully connected route structure.

As hub location is endogenous, asymmetric equilibria where airlines use different hubs may arise. We denote this set of possible equilibria as \((H_x, H_y)\), regardless of the location of the airlines’ hubs. Asymmetric equilibria, with one airline choosing route structure \(u\) and the other \(v \neq u\), will arise if and only if the following holds:

\[
\Pi_i(u, v) \geq \Pi_i(w, v) \land \Pi_i(v, u) \geq \Pi_i(y, u) \quad \forall \ w \in RS, \forall \ y \in RS \ u \neq v .
\] (6.18)

This implies that \(u\) is the best response when the rival chooses \(v\) and vice versa, which again, because of airline, airport and market symmetry, implies that there are multiple equilibria for a particular \(u\) and \(v\). We use \((F, H)\) to denote the asymmetric equilibria where one airline serves the markets with a fully connected route structure and the other with a hub-and-spoke route structure, regardless of the hub airport choice.

Despite having a highly stylized model, these comparisons are hard to perform analytically. To surpass this, we look at some of the relevant equilibrium conditions that, together with numerical examples, allows us to solve the equilibrium, provide intuition and compare our results to those in previous literature.

First, we analyze the expression that characterizes the cases in which using a fully connected structure is a best response to the rival using fully connected as well: \(\Gamma_i = \Pi_i(F, F) - \Pi_i(H_y, F)\). The condition \(\Gamma_i > 0\) is sufficient for \((F, F)\) to be a sub-game equilibrium of the airlines’ game.\(^{72}\) Denote \(f_i(r, v)\) the total number of flights scheduled by airline \(i\) in the third stage equilibrium where it uses \(r\) as a route structure and its rival uses \(v\). Using Eqs. (6.9) and (6.16), \(\Gamma_i\) is:

\[
\Gamma_i = \sum_{m \in M} B \cdot [ (q_i^m(F,F))^2 - (q_i^m(H_y,F))^2 ] + c_f \cdot [ f_i(H_y,F) - f_i(F,F) ] ,
\] (6.19)

where the variables in Eq. (6.19) are evaluated at the untolled equilibrium with route structures indicated in parentheses. Eq. (6.19) shows that there are two types of effects driving the adoption of fully connected over hub-and-spoke: a change in revenues, as a result of the change in the number of passengers in all three markets (first bracketed term in the right-hand side), and a change in costs, due to variations in the total number of flights (second bracketed term in the right-hand side).\(^{73}\) Despite that it is not possible to assess the sign of \(\Gamma_i\) analytically, the sign of each term is intuitive. Hub-and-spoke networks are meant to save airline costs through a reduced number of links flown, thus, a natural expectation is that the total number of flights is reduced when moving from fully connected to hub-and-spoke (Brueckner, 2004). It is, therefore, expected that the second bracketed term is negative, because the total frequency is lower under \(H_y\) than under \(F\), so that this favors the adoption of a hub-and-spoke route structure over a fully connected.

The change in number of passengers in each market, however, may favor the point-to-point structure. To see this, note that the connecting passengers (market \(XZ\)) face

\(^{72}\) Again, we use \(Y\) as the reference hub airport for illustration purpose. Because of the airlines’ and the markets’ symmetry, there is no need to analyze the condition \(\Gamma_i > 0\) for all three airports.

\(^{73}\) The decomposition of \(\Gamma_i\) that we use is helpful to analyze the equilibrium but it should be noted that it is not unique.
higher travel costs under $H_Y$ than under $F$, because they incur higher travel delays and the additional cost of connecting (see Eq. (6.11)). As a result, the equilibrium number of passengers in the connecting market (for the given route structure of the rival) should be higher under a fully connected route structure.\(^{74}\) This demand effect is what favors the adoption of $F$. On the other hand, the full price in the markets that are served directly under both route structures (markets $XY$ and $YZ$) may be higher or lower due to two counteracting effects. Under a hub-and-spoke route structure, the frequency in each link is expected to be higher than under a fully connected route structure as more passengers travel between the hub and the other airports. This higher frequency implies higher congestion and decreased schedule delay costs and, therefore, its effect on the full price is,\(^{a priori}\) ambiguous. However, as markets $XY$ and $YZ$ are served directly under both configurations, the demand effect in these markets is expected to be significantly smaller than in the connecting market, and, therefore, it is not expected to be a main driving force of the sign of $\Gamma_i$.

The driving force behind a symmetric fully connected equilibrium $(F,F)$ is the gain from an increased number of passengers in the connecting market when switching from a hub-and-spoke to a fully connected route structure. This demand increase has to be big enough so that the increase in revenues exceeds the increase in costs that results from operating a higher total number of flights. Therefore, it is expected that $(F,F)$ is an equilibrium, for example, when demand is high and relatively sensitive to (full) price changes, which can be represented by a low value of $B$.

If the profit enhancing effect in the connecting market that favors the fully connected structure dominates when the rival’s route structure is hub-and-spoke, but not when it is fully connected ($\Gamma_i < 0$), $(F,H)$ is an equilibrium.

Conversely, when the cost benefits of adopting a hub-and-spoke route structure are dominant, equilibria with both firms using hub-and-spoke exist. This, as explained above, can be expected to happen when demand is relatively low and insensitive to price changes. Which of the two sets of equilibria that involve both firms adopting hub-and-spoke arises, $(H,H)$ or $(H_x,H_y)$, depends on the sign of $\Phi_i \equiv \Pi_i(H_X,H_Y) - \Pi_i(H_Y,H_Y)$:

$$\Phi_i = \sum_{m \in M} B \cdot \left[ (q_{im}^m(H_X,H_Y))^2 - (q_{im}^m(H_Y,H_Y))^2 \right] + c_f \cdot \left[ f_i(H_Y,H_Y) - f_i(H_X,H_Y) \right] . \quad (6.20)$$

Again, despite that it is not possible to assess the sign of $\Phi_i$ analytically, intuition can be provided. First, if demands are not interdependent, which in our model can only take place when airlines offer independent goods ($E = 0$) and there is no congestion, $\Phi_i$ becomes zero. This is because the route structure configuration and the output choice of the rival would not have any effect on own profit, and, as all markets are symmetric, the location of the hub would not affect profit either. In this extreme case, whenever the profit enhancing effects of adopting a hub-and-spoke route structure dominate the loss due to a demand reduction in the connecting market, both $(H,H)$ or $(H_x,H_y)$ are equilibria. It is

\(^{74}\)With increased travel costs due to the connection, the marginal willingness to pay is shifted downwards and the equilibrium quantity is lower.

\(^{75}\)Note that the direct demand system that results from Eq. (6.2) is: $q_i^n = A/(B + E) - \theta_i \cdot B/(B^2 - E^2) + \theta_j \cdot E/(B^2 - E^2)$. Therefore both the intercept as well as the own-price sensitivity increase as $B$ is reduced.
the demand interdependencies, due to congestion or product substitution, that cause the divergence between these cases.

When $\Phi_i$ is positive, $(H_x, H_y)$ arises instead of $(H, H)$ if, in addition, fully connected is not the best response to the rival using hub-and-spoke. We first look at how the number of passengers in each market differs when adopting hub-and-spoke with a different city as the hub compared to using the same airport as the rival (the first bracketed term on the right-hand side of Eq. (6.20)). The market $XZ$ is served directly under $H_X$ and with a connection under $H_Y$. Given that the rival uses $H_Y$, choosing a different airport as the hub necessarily gives a comparative advantage in this market: by offering a direct flight where the rival offers a connecting service, the firm increases its market share and profit, compared to the case where both serve the market with a connection. The opposite holds true in the market $YZ$, which is served directly by the competitor under $H_Y$, but with a connection by firm $i$ when adopting $H_X$. As a result, there is a demand reduction in the own connecting market that favors using the same city as the hub. Finally, the market $XY$ is served by both firms with direct flights under both settings, so it can be expected that the demand change in this market does not play a major role in shaping the best-response. Something similar can be expected for the change in the total number of flights offered, which is the second bracketed term on the right-hand side of Eq. (6.20). Under both hub-and-spoke route structures, the total number of flights should not vary significantly and we can expect that the frequency costs are not a significant determinant of the sign of $\Phi_i$.

It is therefore expected that the key determinant of the existence of the asymmetric hub-and-spoke equilibrium $(H_x, H_y)$ should be whether the benefits from serving directly the market $XZ$ (the rival’s connecting market) exceed the disadvantages of serving market $YZ$ with a connection (where the rival offers a direct flight). Although these two opposite effects may appear to be of the same size because of symmetry, they differ mainly because of congestion effects. Under $(H, H)$ the hub is heavily congested because all flights either take off or land at airport $Y$, whereas under $(H_x, H_y)$ each hub is less congested. Therefore, by serving directly the market $XZ$, the airline $i$’s passengers face a change from a heavily congested connection to a direct service. In the market $YZ$, the passengers face a change from a heavily congested direct service to a connection that is less congested. This suggests that there is more to be gained in the market $XZ$ than what is to be lost in $YZ$. This is what may make using a different airport as the hub to be the best response, and therefore to give rise to the asymmetric hub-and-spoke equilibria $(H_x, H_y)$.

Finally, note that the conditions analyzed above are not necessarily mutually exclusive and multiple equilibria may arise. To complement the analysis above, we solve the equilibrium of the airlines’ game numerically and provide further intuition.

### 6.3.4 Numerical example

In this section we show the equilibria in route structure of the airlines’ game for a range of the own-demand sensitivity parameter, $B$, and all possible values of the ratio of substi-

---

76Note that we assume that the demand function for each OD pair is the same, which basically means that all origins and destinations are of the same size. In reality, cities differ in size, which makes larger cities more attractive as hubs and this would favor a symmetric network.
tutability, $E/B$, for a particular set of values of the other parameters (full details are in Appendix 6.B). The functional forms used are: $g_i(f_i) = \gamma / f_i$ for the schedule delay cost (where $\gamma$ is the disutility of a unit of schedule delay), and $D(F_i + F_j) = \alpha \cdot (F_i + F_j) / K$ for the users’ congestion cost, where $\alpha$ is the value of travel time and $K$ the airport’s capacity. We choose to display our results in terms of those demand parameters, because they capture key characteristics of the markets. The parameter $B$ allows us to control both the market size and the own-price sensitivity, something that, as discussed above, is expected to have a significant impact in the airlines’ equilibrium: a higher value of $B$ implies a lower demand and a lower own-price sensitivity (see footnote 75). The ratio $E/B$ allows us to study the effect of different degrees of substitutability in our results, and compare them with the case where the demand interdependency is only through congestion. By varying the ratio from 0 to 1, we can assess the effect of product substitution on the results, in presence of congestion.

The remaining parameters are set according to the following criteria: (i) the second-order conditions hold in the studied parameter range (the second-order conditions involving the cross derivatives do not hold globally in our problem); (ii) the market demand price elasticities are in the range provided by InterVISTAS (2007) as representative of the airline industry (between $-0.84$ and $-1.96$); (iii) the own-demand price elasticities are in the range $[-0.1, -3.2]$, found by Brons et al. (2002) as representative, when we consider the ratios of substitutability estimated by Perloff et al. (2007) (between 0.4 and 0.8); and (iv) the airlines’ operating cost parameters are the ones estimated by Swan and Adler (2006) for a long-haul aircraft configuration and a reference trip distance of 5000 km.

Figure 6.2 shows the results in terms of route structure equilibria and reveals that, for the chosen parameters, two set of equilibria arise: both airlines using fully connected route structure ($F, F$), and both using hub-and-spoke but in different hubs ($H_x, H_y$); it also reveals regions where both sets of equilibria may arise (the region between the two lines). The numerical analysis confirms the intuition provided above that for relatively low values of $B$, which imply higher demand levels and higher demand sensitivity, $(F, F)$ is an equilibrium. This is where the effect of a decreased number of passengers in the connecting market dominates the cost benefits, when switching from a fully connected to a hub-and-spoke route structure. The opposite holds true for relatively large values of $B$. Our results in Figure 6.2 also show that, when markets and airlines are symmetric, and for the considered parameters, the best response to the rival using a hub-and-spoke route structure, is either adopting a fully connected or a hub-and-spoke route structure, but using a different airport as the hub. That is the reason why asymmetric hub-and-spoke equilibria $(H_x, H_y)$, arise instead of symmetric hub-and-spoke equilibria $(H, H)$. As the cost advantages that adopting hub-and-spoke brings can be exploited in a similar way under symmetric and under asymmetric hub-and-spoke settings, the main difference between both structures comes from the change in the number of passengers in the connecting market. The numerical example suggests, as discussed previously, that the gain from dominating the rival’s connecting market is higher than the loss of being dominated in the own connecting market, when the reference point is a symmetric setting $(H, H)$.

In addition, sensitivity analyses, in Appendix 6.C, show that results are robust in the sense that, when different parameter values are considered, the absolute position and size of the regions in Figure 6.2 change, but the relative positions and the set of
equilibria that arise remain unchanged. It also confirms the previous results regarding the effect of demand and costs in the resulting equilibria (e.g. Flores-Fillol, 2009), which are summarized as:

- A higher value of the demand intercept ($A$), which implies higher demands, favors the adoption of fully connected route structures. This is in line with the intuition provided above; when markets are larger, the potential loss in the connecting market when adopting hub-and-spoke is increased.

- A higher value of the cost per flight ($c_f$) favors the hub-and-spoke route structure. This result follows directly from the analytical comparisons (see Eq. (6.19)), as the frequency cost advantages that using hub-and-spoke structures brings become more significant.

- A higher value of the marginal cost per seat ($c_q$) favors the fully connected route structure. Connecting passengers require an additional seat in two flights, so that the cost per connecting passenger is higher. This enhances the demand effect that favors the fully connected route structure.

- A higher value of the disutility of schedule delay costs ($\gamma$) favors the hub-and-spoke configuration, as the frequency benefits become more relevant. Concentrating passengers by increasing the frequency through hub-and-spoke structures increases the willingness to pay for the two direct flights and therefore the fare that can be charged.
A higher value of travel time ($\alpha$) favors the adoption of fully connected route structures. As hub-and-spoke route structures induce higher frequencies in the hub-to-spoke links and the traffic is more concentrated, it yields increased congestion. When congestion reductions are valued at a higher price, the increased congestion that hub-and-spoke structures bring shifts the passengers’ willingness to pay downwards.

• A higher value of the constant transfer cost ($\mu$) favors the fully connected route structure. When transferring is more costly for passengers, hub-and-spoke route structures become less attractive.

As the purpose of the chapter is to compare the untolled equilibria with the welfare-maximizing situation, and analyze how to enforce it, the next section studies the first-best combination of route structure and output.

### 6.4 Welfare analysis

We first look at the welfare-maximizing output for a given choice of route structure by airlines and derive the tolls that induce that output choice. Thereafter, we study the socially efficient route structure configuration and whether these tolls are sufficient to achieve it as an equilibrium of the full game.

#### 6.4.1 The symmetric fully connected case

In this case, denoted $(F, F)$, both airlines serve the markets with a fully connected route structure. We look at a regulator that maximizes unweighted social surplus, with the toll per-passenger in each market ($\tau^m_i$) and the toll per-flight in each link ($\tau^l_i$) as instruments. Straightforward calculations yield the following expression for social welfare:

$$SW^{(F,F)} = \left[ \sum_{m \in M} \frac{B}{2} \cdot ((q^m_i)^2 + (q^m_j)^2) + E \cdot q^m_i \cdot q^m_j \right] + \left[ \sum_{m \in M} \tau^m_i \cdot q^m_i + \tau^m_j \cdot q^m_j \right] + \left[ \sum_{l \in L} \tau^l_i \cdot f^l_i + \tau^l_j \cdot f^l_j \right],$$

(6.21)

where the first term in brackets is the consumer surplus, the second term is the airlines’ profit, the third term is the revenue from per-passenger tolls, and the fourth term is the revenue from per-flight tolls. The first-order conditions for welfare maximization under fully connected route structures imply the following pricing and frequency setting rules:

$$\frac{\partial SW^{(F,F)}}{\partial q^m_i} = 0 \Rightarrow p^m_i = c_q \ \forall \ m \in M,$$

(6.22)

$$\frac{\partial SW^{(F,F)}}{\partial f^l_i} = 0 \Rightarrow - \sum_{m \in M} (q^m_i + q^m_j) \cdot \frac{\partial D^m}{\partial f^l_i} + q^m_i \cdot \frac{\partial q^m_i}{\partial f^l_i} = c_f.$$

(6.23)
6.4 Welfare analysis

We drop the index on the delay cost function, as it is the same for both airlines since we are looking at a symmetric route structure configuration. Eq. (6.22) states that the fare should equal the marginal cost of a seat in all markets, and Eq. (6.23) states that, in every link, frequency should be set such that the airline’s marginal cost per flight equals the marginal benefit for all passengers. Comparing Eq. (6.7) with Eq. (6.22), and Eq. (6.8) with Eq. (6.23), we derive the tolls that maximize social welfare under \((F,F)\):

\[
\tau_i^m = -q_i^m \cdot B \quad \forall m \in M ,
\]

\[
\tau_i^l = \sum_{m \in M} q_j^m \cdot \frac{\partial D^m}{\partial f_i^l} \quad \forall l \in L .
\]

This is simply a per-passenger subsidy equal to the markup applied in each market, to eliminate the deadweight losses, and a link-specific per-flight toll equal to the uninternalized congestion. This is a traditional result in the airport pricing literature (e.g. Brueckner, 2005).

The two instruments above (Eqs. (6.24) and (6.25)) attain the social optimum, if both airlines exogenously choose fully connected structures. The optimal value for social welfare, in this fixed route structure equilibrium, is:

\[
SW^{(F,F)} = \sum_{m \in M} \frac{B}{2} \cdot ( (q_i^m)^2 + (q_j^m)^2 ) + E \cdot q_i^m \cdot q_j^m - \sum_{l \in L} (f_i^l + f_j^l) \cdot c_f ,
\]

with quantities and frequencies satisfying Eqs. (6.22) and (6.23).

6.4.2 The symmetric hub-and-spoke case

Following the same procedure as in Section 6.4.1, straightforward calculations yield the following rules for welfare-maximizing pricing and frequency setting under the \((H,H)\) route structure (with \(Y\) as the hub airport in this case):

\[
\frac{\partial SW^{(H,H)}}{\partial q_i^m} = 0 \Rightarrow p_i^m = c_q \quad \forall m \in \{XY,YZ\} ,
\]

\[
\frac{\partial SW^{(H,H)}}{\partial q_i^{XZ}} = 0 \Rightarrow p_i^{XZ} = 2 \cdot c_q ,
\]

\[
\frac{\partial SW^{(H,H)}}{\partial f_i^l} = 0 \Rightarrow - \sum_{m \in \{XY,YZ\}} (q_i^m + q_i^{XZ} + q_j^m + q_j^{XZ}) \cdot \frac{\partial D^m}{\partial f_i^l} - (q_i^m + q_i^{XZ}) \cdot \frac{\partial q_i^m}{\partial f_i^l} = c_f .
\]

Again, in the optimum, the fare should equal the marginal cost in all markets, and frequencies should be set such that the airline’s marginal cost per flight equals the marginal benefit for all passengers. Comparing first-order conditions, we obtain the tolls that maximize social welfare under symmetric hub-and-spoke route structures:

\[
\tau_i^m = -q_i^m \cdot B \quad \forall m \in M ,
\]
These are the sufficient instruments when route structure is fixed to be hub-and-spoke for both airlines. In equilibrium, the optimal value for social welfare under \((H, H)\) can be written as:

\[
SW^{(H, H)} = \sum_{m \in M} \frac{B}{2} \cdot [q_m^{(F)}]^2 + \frac{E}{2} \cdot q_i^m \cdot q_j^m - \sum_{l \in \{xy, yz\}} (f_l + f_j) \cdot c_f, \tag{6.32}
\]

with variables satisfying Eqs. (6.27)–(6.29).

### 6.4.3 The asymmetric cases

We have shown in Section 6.3.3 that also asymmetric equilibria may arise, in particular \((F, H)\) and \((H_x, H_y)\). It is straightforward to show that the optimal pricing and frequency setting rules will be a combination of the rules described above (Eqs. (6.22), (6.23), and (6.27)–(6.29)). As a result, the tolls that maximize welfare for asymmetric equilibria will also be a combination of the tolls above; for an airline with fully connected route structure, the tolls in Eqs. (6.24) and (6.25) should be charged, and for an airline using hub-and-spoke, the ones in Eqs. (6.30) and (6.31). The difference will be that the tolling rules will be evaluated at different outputs, and that Eqs. (6.30) and (6.31) have to be adjusted if the hub is not \(Y\).

### 6.4.4 The optimal route structure

We now look at the combination of route structure and output that maximizes social welfare. Again, complexity prevents us from fully comparing social welfare values analytically. As in Section 6.3, we combine analytical results with numerical examples to identify the equilibria and provide intuition. We also compare the choice of route structure by unregulated firms with the welfare-maximizing choice, to identify the potential sources of inefficiency.

Consider first the comparison between the following route structure configurations: \((F, F)\) and \((H, H)\). Let the difference between social welfare in both settings be \(\Delta \equiv SW^{(F, F)} - SW^{(H_y, H_y)}\), where the airport \(Y\) is used as the hub for illustration purpose. The condition \(\Delta > 0\) is necessary for \((F, F)\) to be the welfare-maximizing route structure setting. Using Eq. (6.26), Eq. (6.32), and symmetry of the first-best solution we get:

\[
\Delta = \sum_{m \in M} (B + E) \cdot [\left(q_m^{(F)}\right)^2 - \left(q_m^{(H_y, H_y)}\right)^2] + 2 \cdot c_f \cdot [f_{(H_y, H_y)} - f_{(F, F)}], \tag{6.33}
\]

where the variables in Eq. (6.33) are defined as \(q^m = q_i^m = q_j^m \forall m \in M\) and \(f = f_i = f_j\) because of symmetry, and they are evaluated at the social optimum for the given route structure setting in parentheses. The comparison between \(\Gamma_i\) in Eq. (6.19), which defines the sufficient condition for \((F, F)\) to be an equilibrium \((\Gamma_i > 0)\), and \(\Delta\) in Eq. (6.33) sheds light on the inefficiency of route structures choice by profit maximizing agents. It
also makes implausible that unregulated competition between airlines will always lead to
the first-best route structure. First, recall that there is a difference in output between the
two cases, as discussed above, due to two effects: market power exertion and the presence
of congestion externalities. This clearly makes the variables in ∆ and Γi to differ. Even if
the outputs were the same, in the social welfare comparison there is a term involving the
cross sensitivity parameter (E) that is absent in profit comparison; i.e. a firm ignores the
effect of its choices on the consumer surplus derived by the competitor’s passengers. In
addition, in the welfare comparison (∆), the cost difference due to a different total number
of flights takes into account the savings for both firms (the factor of 2 multiplying the
last term on the right-hand side of Eq. (6.33)). Finally, the airlines’ relevant comparison
is for a given route structure of the rival, which again makes ∆ and Γi diverge.

A look at ∆ in Eq. (6.33) reveals that the driving forces behind which route structure
composition maximizes welfare, between the symmetric fully connected (F, F) and the
symmetric hub-and-spoke (H, H) settings, are similar to the ones driving the airlines’
choice. This is, there is a cost effect (second term on the right-hand side of Eq. (6.33))
that favors the hub-and-spoke over the fully connected route structure if the total number
of flights is lower, which is a natural expectation. What favors the fully connected route
structure is a higher consumer surplus in the market that is served with a connection under
(H, H). When market XZ is served directly, the social cost (hence the full price paid) is
lower because there are no connection costs and because a passenger requires an additional
seat in one link instead of two (see Eqs. (6.27) and (6.28)). Therefore, demand is higher in
that market when served directly and the consumer surplus increases. Again, as a result,
we can expect that (F, F) yields higher welfare than (H, H) when demand is relatively
high and price sensitive (low values of B). When this is not the case, and the cost benefits
that hub-and-spoke route structures bring exceed the drawbacks from having connecting
passengers, settings involving hub-and-spoke will be welfare maximizing. To gain insights
on which of those settings could maximize welfare, let \( \Lambda \equiv SW(H_X,H_Y) - SW(H_Y,H_Y) \) be
the difference between social welfare under an asymmetric hub-and-spoke setting and a
symmetric configuration involving hub-and-spoke. Straightforward calculations lead to:\n
\[
\Lambda = \sum_{r \in \{i,j\}} c_f \left[ f_r(H_Y,H_Y) - f_r(H_X,H_Y) \right] + \sum_{m \in \mathcal{M}} \sum_{r \in \{i,j\}} B \left[ \left( q_r^m(H_X,H_Y) \right)^2 - \left( q_r^m(H_Y,H_Y) \right)^2 \right] + \sum_{m \in \mathcal{M}} E \left[ q_r^m(H_X,H_Y) \cdot q_r^m(H_X,H_Y) - q_r^m(H_Y,H_Y) \cdot q_r^m(H_Y,H_Y) \right],
\]

(6.34)

where the variables in Eq. (6.34) are evaluated at the social optimum for the given route
structure setting in parentheses, in which the first structure refers to firm i and the second
to firm j. What drives which of these two settings yields higher welfare is again related to
the number of passengers who face a higher social cost and the difference in the airlines’
costs. However, as both firms use a hub-and-spoke route structure in both configurations,
the economies of density are exploited in similar ways and the difference in costs should not
be significant. In other words, the total number of flights should not differ much between
the two configurations, and, therefore, the first term on the right-hand side of Eq. (6.34)
should not play a major role in determining the socially optimal setting between these
two configurations.
Before elaborating about the sign of each of the remaining terms, note that if products are independent (i.e. \( E = 0 \)) and there is no congestion, \( \Lambda = 0 \) holds and both settings yield exactly the same welfare. This is because demands are no longer interdependent and, as all markets are symmetric, the hub location would not have any effect on welfare. Consider now that the products are independent, but there is congestion. In this case, the difference between the two settings, \((H_X, H_Y)\) and \((H_Y, H_Y)\), is only due to differences in congestion. In the symmetric case, the hub (\(Y\) in this example) is congested by the two services of both airlines, whereas in the asymmetric case the airports that are used as a hub are congested by two services of one airline and one service of the other. This should result in a less congested situation and it should favor the asymmetric setting.

When there is also imperfect substitution, the number of connecting passengers, who are the ones facing a higher full price, is lower under \((H_X, H_Y)\). When firms use hub-and-spoke with different airports as their hub, in each market that is served with a connection, one airline provides a direct service at marginal social cost, which is lower than the marginal social cost of the connection. Therefore, the effect of imperfect substitution also favors the asymmetric over the symmetric hub-and-spoke setting. In the limit, when products are perfect substitutes, the social optimum cannot have two airlines using different route structures (e.g. \((H_X, H_Y)\)), because full prices must be the same for all airlines that are serving the market, but also should be set at marginal social cost. As marginal social costs, under different route structures, are different in at least one market, these two constraints make an asymmetric route structure setting incompatible with welfare maximization when products are perfect substitutes \((E/B = 1)\). What is optimal under perfect substitution, instead, is to have regulated monopolized markets. This is one airline serving all three markets either with a fully connected \((M_F\) hereafter) or with a hub-and-spoke route structure \((M_H\) hereafter). The reason behind this result is that higher welfare is achieved under the full regulation of a monopoly than of perfect substitute competing airlines. This is because we are looking at a regulator who solves the congestion inefficiency and the market power exertion through tolls; therefore, there is no deadweight loss regardless of the number of firms. In addition, the frequency set by a fully regulated monopoly airline is higher than for a fully regulated airline in oligopoly, as demand is divided between firms, thus the schedule delay costs will be lower. This is what favors a monopoly. When differentiation is weak \((B \approx E)\), this effect of lower schedule delay costs may still be dominant, and, as a result, the regulation of a monopoly may still yield higher welfare than of a duopoly. When differentiation is strong \((B \gg E)\), the expansion of demand generated by a new firm may overweight the schedule delay cost advantages of a monopoly, and competition will bring higher welfare.\(^{77}\)

Figure 6.3 summarizes, for the same parameter region used in Section 6.3 (see Appendix 6.B for details), the welfare-maximizing route structure (when evaluated at the optimal output for those parameters). Recall that the first-best setting in Figure 6.3 is the one that maximizes welfare, and it is not necessarily an equilibrium of the tolled competition. We analyze whether and when the first-best setting and the (tolled) equilibrium coincide in the following section.

Figure 6.3 suggests that, for the chosen parametrization, the route structure config-

\(^{77}\)Despite having a different model, the intuition is the same as in Basso (2008).
6.4 Welfare analysis

Figure 6.3: Welfare-maximizing route structures. Main parameterization (see Appendix 6.B).

Figure 6.3 also suggests that the higher the substitutability between airlines (higher $E/B$), the more likely is $(H_x, H_y)$ to be a more efficient route structure configuration than $(F, F)$. It still brings cost savings and frequency benefits, but it is less essential that all airlines are present on all routes. As explained above, one of the effects that favors the fully connected symmetric setting over the hub-and-spoke structures is the lack of connecting passengers, who have a higher marginal social cost. In the asymmetric hub-and-spoke setting $(H_x, H_y)$, the number of connecting passengers decreases as the airlines are perceived as closer substitutes, because in every connecting market of one airline, the rival provides a direct service priced at marginal social cost (because we are looking at the first-best setting). Therefore, when products are close substitutes, the number of connecting passengers is low, and the gains from lower total costs due to reduced total frequency of $(H_x, H_y)$ dominate the possible gains due to higher aggregate demand in the connecting markets under $(F, F)$.

As Figure 6.3 reveals, the result that monopolized markets yield higher welfare also holds when airlines are close substitutes: in the parameter region $M_f$, it is welfare maximizing to have a regulated monopoly airline serving the three markets with a fully connected route structure.

We now turn to the analysis of how to enforce the first-best described in this section.
That is, can the first-best setting be a toll-decentralized equilibrium?

### 6.5 Sufficient instruments for social welfare maximization

#### 6.5.1 First-best analysis

In order to study whether the two pricing instruments described in Section 6.4 align airline choices with welfare maximization, we numerically examine the equilibrium of the game when the regulator charges the optimal tolls conditional on the first-best route structure. In other words, we derive the outcome of the game in each of the regions of Figure 6.3, when the regulator charges the tolls that induce the optimal output for the given welfare-maximizing route structure. For example, in the parameter region denoted by \((F, F)\) in Figure 6.3, the regulator set the tolls according to rules in Eqs. (6.22) and (6.23); if the equilibrium that results from charging these tolls is with both airlines choosing the fully connected route structure, the first-best is decentralized as the charges ensure optimal outputs. Conversely, if the equilibrium with the optimal charges in the parameter region where \((F, F)\) maximizes welfare is not with both airlines choosing the fully connected route structure, we can conclude that the two pricing instruments are not sufficient in this case. This is because any other charge, that may induce the optimal route structure configuration as an equilibrium, will not induce the optimal output.

Let us first discuss the difference between the route structure in the untolled equilibrium versus that in the optimum (later we will discuss whether tolls alone can decentralize the optimum). Figure 6.4a compares the untolled equilibrium in Figure 6.2 with the welfare-maximizing setting in Figure 6.3, in terms of route structure. It reveals that the rationale behind the charges is not always the same: they can be required only to correct output setting, to correct simultaneously output and route structure choice, and in order to correct market structure as well. The white areas represent the cases where the airlines' route structure equilibrium is unique and the same as the first-best, and only output corrections are needed. The light gray areas show where there is a route structure equilibrium of the untolled competition that is different from the efficient one, and the tolls are required to induce airlines to choose both the welfare-maximizing outputs and route structures. The \(M_f\) region, in dark gray, indicates that welfare maximization requires tolls that exclude one airline from the market, as a fully regulated monopoly would be optimal. We use the labels to indicate the efficient route structure configuration when it is not the unique outcome of the unregulated competition.

Figure 6.4a also shows that the result that a monopoly airline exhibits a bias towards hub-and-spoke route structure does not necessarily carry over to competing airlines.\(^78\) That is, no longer whenever the welfare-maximizing setting has firms using hub-and-spoke structures, it is also an equilibrium of the untolled competition. For the chosen

\(^78\)In a monopoly context, Brueckner (2004) shows that the choice of route structure by a monopoly airline is biased towards hub-and-spoke. In our problem, the divergence between settings is more complex because frequency setting—for a given number of passengers—is distorted by congestion effects, and, when airlines are substitutes, both are distorted by strategic effects.
Figure 6.4: First-best tolls’ rationale (6.4a) and equilibrium under optimal pricing (6.4b). In the shaded areas of Figure 6.4a there are equilibria of the untolled competition that are different from the welfare-maximizing route structure configuration and labels indicate the welfare-maximizing route structure configuration. Labels in Figure 6.4b indicate the equilibrium in route structures under optimal airport pricing, and the shaded areas highlight the cases where tolls cannot always decentralize the first-best. Main parameterization (see Appendix 6.B).

parametrization, \((H_x, H_y)\) is optimal but the untolled equilibrium is \((F, F)\) when the demand, the demand sensitivity to own-price changes and the degree of substitutability are not too low (roughly, in the triangle where \(B\) is lower than 40 and \(E/B\) between 0.15 and 0.9 in Figure 6.4a). Recall that in that area we find \((F, F)\) in the untolled equilibrium because, when demand is relatively high and sensitive to price changes (low \(B\)), the profit loss in the connecting market is larger than the cost benefits that may be obtained when switching from a fully connected to a hub-and-spoke route structure. However, the demand reduction in the airlines’ connecting market does not necessarily harm welfare. In this area, we find \((H_x, H_y)\) in the optimum because the cost advantages that hub-and-spoke structures bring can be exploited without having a large number of connecting passengers, who are the ones with a higher social cost. Under \((H_x, H_y)\), in each airline’s connecting market the competitor offers a direct service with a lower full price. As demand is relatively sensitive to price changes and the substitutability between firms is not too low, the resulting number of connecting passengers is low. This is only possible with airlines adopting different airports as their hub, as otherwise the number of connecting passengers would not necessarily decrease.

The results also indicate that a “naive” regulator, who observes the unregulated equilibrium and sets the tolls based on the then observed route structure, may not always achieve the first-best. The regulator should realize whenever the observed equilibrium is not efficient in terms of route structure (the light gray regions), and induce airlines, via tolls, to change the way they serve the markets. For example, a regulator that plans to
put congestion pricing into action may need to set optimal toll levels that are far from what would be optimal if the route structure choice was exogenous.

Figure 6.4b shows the subgame-perfect Nash equilibrium in terms of route structure, i.e. the equilibrium in route structure choice when the regulator applies optimal pricing designed to induce the optimal output based on the welfare-maximizing route structure configuration. A main result of our numerical analyses is that the first-best cannot always be enforced by using the airline- and market-specific per-passenger tolls together with the airline- and link-specific per-flight tolls designed to induce the optimal outputs of the first-best route structures. Thus, the subgame-perfect Nash equilibrium is not always efficient, even when the regulator can perfectly discriminate between airlines and when it has no budget constraints. The shaded areas in Figure 6.4b represent the parameter range where the toll instruments described in the previous sections cannot always decentralize the first-best. In these regions, airlines adopting fully connected route structures is an equilibrium whereas the first-best setting is \((H_x, H_y)\). As Figure 6.4b shows, there is a region in which \((H_x, H_y)\) is welfare maximizing but the unique equilibrium from the tolled competition is \((F, F)\), and there is also a region in which \((H_x, H_y)\) is welfare maximizing but \((F, F)\) is one of the two possible equilibria.

To provide the intuition behind this result, let us focus on firm \(i\). The welfare-maximizing setting in the shaded areas in Figure 6.4b is \((H_x, H_y)\) and it has firm \(i\) offering a connection in a market \((YZ)\) in which the rival offers a direct service with a lower full price. This maximizes welfare because it yields a low number of connecting passengers (who face a higher social cost) when the firms’ products are imperfect substitutes. However, while having firm \(i\) offering a connection and having a small market share in the market \((YZ)\) is welfare enhancing, it is, at the same time, profit reducing. From the airline \(i\) point of view, given that the competitor flies directly in that market and substitutability is not low, flying directly may be the best response, as it increases the willingness to pay of own-passengers (no layover time cost) and reduces the marginal cost (one seat instead of two). This is what occurs in the shaded regions of Figure 6.4b, where the misalignment of the market \((YZ)\)’s demand effect on welfare and profit makes the route structure equilibrium to be different than the optimal configuration. To ensure that the first-best outcome is reached in these cases, an additional instrument is therefore required.\(^{79}\)

Another policy implication of our analysis is that more information is needed to attain the first-best, as the variables usually needed for setting first-best tolls (e.g. marginal external cost) are not enough to assess discrete changes in welfare under different route structures. A question that naturally follows is how does pricing perform when the optimal network configuration is not assessed or known by the regulator. This is, what is the relative efficiency of setting the tolls based on the observed unregulated route structure equilibrium (in Figure 6.2). This naive pricing is inefficient whenever the untolled equilibrium differs in terms of route structure from the welfare-maximizing setting (the light

\(^{79}\)For example, an arbitrarily high per-flight toll to each airline only when they fly directly in the market that is meant to be served with a connection can act as a barrier to fully connected structures. As this charge is not paid in equilibrium, because the airlines adopt a hub-and-spoke structure, it does not induce any inefficiency in output, but only the incentives to adopt the optimal route structure. A direct restriction to an airline to fly from those airports is also a sufficient additional instrument for these cases.
6.5 Sufficient instruments for social welfare maximization

gray areas of Figure 6.4a). Figure 6.5a shows the relative efficiency achieved by a “naive” regulator, ω, as the percentage of the maximum welfare that can be gained (relative to the untolled equilibrium as the reference scenario). It reaches a minimum of 0.9 in the darkest gray regions of Figure 6.5a and a maximum of 1 in the white areas. We assume that the inefficient setting arises as equilibrium in the cases where multiple equilibria may arise, in order to assess the minimum value of ω without using an equilibrium selection mechanism.

Figure 6.5: Relative efficiency of naive airport pricing based on the untolled route structure equilibrium (6.5a) and relative efficiency of optimal airport pricing (6.5b). Main parameterization (see Appendix 6.B).

Another question that follows from the above results is how big the loss in welfare is, when charging the tolls that correct the output inefficiencies and not achieving the first-best route structure. Again, in this exercise, we look at the case where the inefficient route structure equilibrium arises in the region where multiple equilibria may arise (see Figure 6.4b), to assess the worst-case scenario without analyzing an equilibrium selection mechanism. Therefore, there will be a welfare loss where the optimal tolls cannot always decentralize the first-best in terms of route structure, the shaded areas of Figure 6.4b. There, \((H_x, H_y)\) is the optimal route structure configuration, but \((F, F)\) arises as an equilibrium under optimal pricing and, therefore, tolls are designed to induce the optimal output for the suboptimal \((F, F)\) configuration. Figure 6.5b shows the relative efficiency of such tolls (ω), as the percentage of the maximum welfare that can be gained. The relative efficiency ranges from 1, in the white areas of Figure 6.5b where the tolls are sufficient instruments to achieve the first-best, to 0.99, in the darkest region of the figure. This suggests that although there is a theoretical possibility of decentralizing the “wrong” first-best route structure configuration, the welfare losses may be modest.

Although, in our model, the regulator does not have direct control on the number of airlines, he can price an airline out of one market through the tolls. To do this, one
Chapter 6 Airline route structure competition and network policy

airline should receive the per-passenger (market power) subsidy that corresponds to the monopoly output, and face no congestion toll. This is because a monopolist perfectly internalizes congestion externalities. The other airline should face a prohibitive toll that removes the incentives to become active. Note that the latter may not always be needed; in some cases, a large output choice by the heavily subsidized airline may be enough to shift the rival’s demand to levels that are not profitable or non-positive.

Finally, the sensitivity analyses in Appendix 6.C suggest that the results are robust in that only the boundaries indicated in Figure 6.4 change when parameters are changed, but the main qualitative results hold. Specifically, we find that the tolls are not always sufficient to achieve the first-best when varying the cost parameters ($c_f$ and $c_q$), the demand intercept ($A$), the disutility of schedule delay costs ($\gamma$), and the congestion levels (e.g. through different airport capacities). Also in those cases, the relative efficiency of optimal airport pricing ($\omega$) and naive airport pricing ($\bar{\omega}$) remains high.

6.5.2 Second-best analysis

The results in Section 6.5.1 require the regulator to be able to give subsidies to airlines, and, in some cases, to exclude one airline out of the market. This is arguably close to impossible to carry out in real networks, and, instead, a regulator will most likely be constrained to charge non-negative tolls. This section considers the more realistic case where only non-negative tolls can be charged, and compares the second-best optimal route structure configuration with the first-best setting. An important policy question is, again, whether the second-best tolls are able to induce the desired (second-best) outcome also in terms of route structure.

As we normalize airport costs to zero, the first-best toll per passenger we found above is equal to a market power subsidy, and therefore it is always negative (see Eqs. (6.24) and (6.30)). The first-best per-flight toll is always non-negative (un-internalized congestion, see Eqs. (6.25) and (6.31)). It is, therefore, expected that in the second-best case we now study, the per-flight tolls are adjusted downwards to compensate for the lack of subsidies, and that in some parameter region they may even become (constrained to) zero. This would be the case if the inefficiency from the congestion externality is sufficiently low compared to the inefficiency from the market power exertion.\(^{80}\)

Figure 6.6 shows the second-best optimal setting for the same parameter constellation and functional forms as in previous sections. The solid line divides the different second-best optimal regions, which are again $(F, F)$ or $(H_x, H_y)$, and the dashed line is the analogous divisional line for the first-best case. Note that we omit, in Figure 6.6, the region in which regulating monopolized markets is optimal when the regulator can subsidize airlines (see Figure 6.3). This allows for a comparison between the second- and first-best optimal route structure configuration.

Figure 6.6 shows that also in the second-best situation considered, $(F, F)$ is the welfare-maximizing setting for low values of $B$ (relatively high demand and high price sensitivity), and $(H_x, H_y)$ for higher values of $B$. In a similar way as in the first-best case, a higher ratio of substitutability favors the asymmetric hub-and-spoke over the symmetric fully

---

\(^{80}\)Pels and Verhoef (2004) analyze this using a single origin-destination pair. They find cases where the second-best policy is to set the tolls equal to zero.
6.5 Sufficient instruments for social welfare maximization

Figure 6.6: Second-best optimal route structures (solid line) and comparison with the first-best case (dashed line). Main parameterization (see Appendix 6.B).

Despite the similarity, there is a region —denoted \( P \)— where \((H_x, H_y)\) is second-best optimal whereas \((F, F)\) is first-best optimal, and a region —denoted \( O \)— where \((F, F)\) is second-best optimal whereas either \((H_x, H_y)\) or \((M_f)\) is first-best optimal. The difference is caused by that in the second-best cases, there are deadweight losses due to the lack of subsidies and the demand is lower than in the first-best. The monopoly region does not exist for the considered parameters, as the monopolist’s market power exertion is no longer being corrected through the subsidy; it is better to regulate a duopoly when tolls are constrained to be non-negative. When both airlines use fully connected, competition is direct in all three markets, whereas in the asymmetric hub-and-spoke equilibrium there are two markets where one airline has an advantage over the other. Therefore, the direct competition that a fully connected route structure brings may decrease deadweight losses by more than under an asymmetric hub-and-spoke structure. This effect, absent in the first-best case, favors \((F, F)\). On the other hand, as demand is lower, frequencies are lower and the cost advantages of hub-and-spoke structures are more attractive. Phrased differently, there are more incentives to exploit the economies of density in the second-best that in the first-best case. In the area \( O \), substitutability \((E/B)\) is high and demand is high and more price sensitive (low \( B \)), so that the effect of the reduction of the deadweight losses dominates, and the more intense competition in \((F, F)\) yields higher welfare in the second-best, whereas \((H_x, H_y)\) or \((M_f)\), where subsidies directly stimulate demand, is first-best optimal. In the area \( P \), where demand is lower and less price sensitive (higher
and substitutability \((E/B)\) is not too high, these effects are reversed, and the cost advantages of hub-and-spoke structures become more attractive, making \((H_x, H_y)\) second-best optimal where \((F, F)\) is first-best optimal, given the higher demand that the first-best subsidies induce.

Figure 6.7a compares the second-best optimal route structure configuration with the untolled equilibria, and it reveals that also the rationale for the second-best charges may include the desire to change the route structure. The white areas again indicate the cases where the route structure in the untolled equilibrium is unique and the same as in the second-best optimal setting, and tolls are only needed to correct for output inefficiencies. The gray region shows where there is an untolled equilibrium in route structures that is different from the second-best optimal configuration, and therefore tolls ideally would be required, but sometimes are unable, to correct both. The label indicates the second-best optimal route structure setting when an inefficient untolled equilibrium exists.

Finally, where the second-best optimal route structures in Figure 6.6 are assumed to apply, we also assess whether they are decentralized by non-negative tolls, in order to study whether the per-flight tolls are sufficient instruments to achieve the second-best outcome in route structures shown in Figure 6.6. The main result of this exercise is that a regulatory instrument designed to correct route structure may be needed in addition to the per-flight tolls: there are different regions where the second-best optimal route structures and outputs differ from the airlines’ equilibrium with non-negative charges. Figure 6.7b summarizes these results by showing the equilibria of the tolled competition.
The gray regions show the cases where the second-best optimal setting \((H_x, H_y)\) cannot always be enforced by a regulator constrained to non-negative tolls, and without additional instruments. Although the constrained per-flight tolls are positive in these regions, they are not always able to decentralize the second-best optimal configuration. In these gray areas, \((H_x, H_y)\) is the second-best optimal configuration, but both airlines using fully connected \((F, F)\) is an equilibrium when they face the tolls designed to induce the second-best optimal output under \((H_x, H_y)\). Again, there is a region where \((F, F)\) is the unique equilibrium under non-negative tolls whereas \((H_x, H_y)\) is second-best optimal, and a region where \((F, F)\) is one of the two possible equilibria under non-negative tolls whereas only \((H_x, H_y)\) is second-best optimal.

This result extends our finding that optimal pricing may not decentralize the choice of the optimal route structure configuration to the second-best case. However, when subsidies are not feasible, we find that the potential welfare gains from second-best tolling are limited. As a consequence, the welfare losses of not having the second-best optimal route structure configuration are low. Therefore, also in the second-best case, our results suggest that there is a chance of decentralizing a suboptimal route structure configuration through optimal pricing, but that the welfare losses may be moderate.

6.6 Conclusions

In the present chapter, we have compared the unregulated equilibrium of route structure competition between airlines with market power, in the presence of congestion, with the welfare-maximizing setting. A first main finding from our analysis is that the unregulated equilibrium in route structure may be different than the welfare-maximizing one, and that a regulator may not always be able to decentralize the first-best route structures using the per-passenger tolls and the per-flight tolls that correct output inefficiency, even if these tolls are allowed to be market and airline specific. Secondly, we also study the case where a regulator is constrained to use second-best non-negative tolls. Again, the second-best optimal route structure may be different from the outcome of unregulated competition, and again, when the tolls are constrained to be non-negative, the second-best optimal setting may not be decentralized with tolls.

An important policy implication from our analysis is that a regulator may benefit from using an additional regulatory instrument, different from the tolls that are only concerned with the congestion externality and the market power exertion. The additional instrument has to align the airline choice of route structure with the welfare-maximizing one, without affecting the choice of output. For example, a direct restriction on the links that can be flown by each airline can solve the problem, or an arbitrarily high toll in the links that are not supposed to be used in equilibrium can also induce the first-best setting.

Another policy implication from our results is that regulators designing pricing policies need more information than what would be required if route structures were exogenous. Even if the policymaker cannot use an additional regulatory instrument and wants to, for example, put congestion pricing in place, it needs to assess discrete changes in welfare due to discrete changes in route structures. Marginal benefits and marginal costs are not

\[81\text{In the studied parameter range, the potential gains from second-best tolling are not higher that } 3\%.\]
always enough information to design a pricing policy. It may happen that in order to fully reap the long-run benefits from congestion pricing, the tolls should be designed for a different route structure configuration than the one that is observed. Nevertheless, our results suggest that the relative welfare losses from inducing the wrong network configuration through tolling, be it first-best or second-best, are limited. So, while the conceptual point that optimal pricing may not decentralize the choice of optimal route structures is significant, the consequent welfare losses may be limited. The latter result, in itself, is of course also an important insight for practical policy making. In any case, it appears to be so in the small numerical example we have developed and its real relevance is a matter of empirical investigation.

All of the results above, as additional sensitivity analyses show, seem robust to parameter values. Still, for our analysis we have used the simplest possible model that allows us to study optimal pricing in networks. We see extending the model as a natural avenue for future research. For example, the consideration of a larger network (more nodes) and the interaction between several airlines are logical extensions; the role of the endogenous hub location is likely to be important in those settings. Considering asymmetry of markets and airlines is also an important topic for future research, as the competition between regional, national, and low-cost carriers is one of the driving forces of route structure adoption in aviation networks. Finally, this framework can be extended to analyze how airports regulated by different authorities interact and affect route structure equilibrium, as well as to the welfare implications of alliances and merges.
Appendix 6.A  Glossary of notation

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>Inverse demand intercept (reservation price)</td>
</tr>
<tr>
<td>$B$</td>
<td>Inverse demand own-quantity sensitivity parameter</td>
</tr>
<tr>
<td>$c_f$</td>
<td>Fixed operating cost per flight of firm $i$</td>
</tr>
<tr>
<td>$c_q$</td>
<td>Operating cost per seat per flight of firm $i$</td>
</tr>
<tr>
<td>$D_i^m$</td>
<td>Passengers’ cost of delays when traveling with airline $i$ in market $m$</td>
</tr>
<tr>
<td>$E$</td>
<td>Inverse demand cross-quantity sensitivity parameter</td>
</tr>
<tr>
<td>$E/B$</td>
<td>Degree of substitutability between airlines’ product</td>
</tr>
<tr>
<td>$F$</td>
<td>Fully connected route structure</td>
</tr>
<tr>
<td>$f_l$</td>
<td>Flight frequency of the airline $i$ in link $l$</td>
</tr>
<tr>
<td>$g_i^m$</td>
<td>Schedule delay cost for a passenger traveling with airline $i$ in market $m$</td>
</tr>
<tr>
<td>$H$</td>
<td>Hub-and-spoke route structure</td>
</tr>
<tr>
<td>$H_C$</td>
<td>Hub-and-spoke route structure using airport $C$ as the hub</td>
</tr>
<tr>
<td>$K$</td>
<td>Capacity of the airport</td>
</tr>
<tr>
<td>$p_i^m$</td>
<td>Fare charged by airline $i$ in market $m$</td>
</tr>
<tr>
<td>$q_i^m$</td>
<td>Number of passengers traveling with airline $i$ in market $m$</td>
</tr>
<tr>
<td>$s_l$</td>
<td>Aircraft size of firm $i$ in link $l$</td>
</tr>
<tr>
<td>$(F,F)$</td>
<td>Route structure configuration with both firms using fully connected</td>
</tr>
<tr>
<td>$(H_e,H_y)$</td>
<td>Route structure configuration with both firms using hub-and-spoke</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Passengers’ value of travel time</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>Passengers’ willingness to pay to reduce a unit of schedule delay time</td>
</tr>
<tr>
<td>$\theta_i^m$</td>
<td>Full price for a passenger traveling with airline $i$ in market $m$</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Additional cost of layover time of a connecting flight</td>
</tr>
<tr>
<td>$\tau_l$</td>
<td>Per-flight toll charged to firm $i$ in link $l$</td>
</tr>
<tr>
<td>$\tau_i^m$</td>
<td>Per-passenger toll charged to firm $i$ in market $m$</td>
</tr>
</tbody>
</table>

Table 6.1: Glossary of notation.

Appendix 6.B  Functional forms and parameters of the main case

We use the following functional forms for the schedule delay cost ($g_i$) and passengers’ congestion cost ($D$):

$$g_i(f_i) = \gamma \cdot \frac{1}{f_i}, \quad D(F_i + F_j, K) = \alpha \cdot \left(\frac{F_i + F_j}{K}\right),$$

where the schedule delay cost in (6.35) is inversely proportional to the airline frequency, and $\gamma$ is a constant representing the monetary value of a unit of schedule delay time. The
congestion delay at each airport in (6.35) is a function of the volume capacity ratio, with \( \alpha \) being proportional to the passengers’ value of travel time, and \( K \) the capacity. The parameter values of our main application are summarized in Table 6.2:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>A</th>
<th>B</th>
<th>E/B</th>
<th>( \alpha )</th>
<th>( \gamma )</th>
<th>( \mu )</th>
<th>( K )</th>
<th>( c_q )</th>
<th>( c_f )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>4400</td>
<td>( \in [15,50] )</td>
<td>( \in [0,1] )</td>
<td>128,25</td>
<td>80</td>
<td>40</td>
<td>4.85</td>
<td>82.8</td>
<td>17470.8</td>
</tr>
</tbody>
</table>

Table 6.2: Parameter values.

The airlines’ cost parameters are from Swan and Adler (2006) for a 5000 km. trip and a long-haul aircraft configuration. The remaining parameters are set to obtain realistic values of elasticities and aircraft size. Perloff et al. (2007) estimate linear demand functions for two Chicago-based markets in which two airlines had a joint share of 98% of passengers, so that it can be regarded as a case of duopolistic competition. They find ratios of substitutability between 0.4 and 0.8. With our parameters in the untolled competition, and with a ratio of substitutability of 0.8, the minimum elasticity in a direct market under \((F,F)\) is -2.94 and under \((H_x,H_y)\) is -3.12. When \( E/B \) is 0.4, we find a maximum elasticity in a direct market under \((F,F)\) of -1.25 and under \((H_x,H_y)\) is -1.31. All values are in the range of \([-0.1, -3.2]\) found by Brons et al. (2002) in their meta-analysis of elasticities of the airline industry.

The market elasticities that we find under \((F,F)\) and under \((H_x,H_y)\) are in the range \([-0.86, -2.08]\), which is roughly the same as the one found by InterVISTAS (2007): \([-0.84, -1.96]\). Finally, in the equilibrium of the untolled competition we find a minimum aircraft size of 70 seats and a maximum of 232 seats.

Appendix 6.C  Sensitivity analysis

In this section, we show how our results change when different parameter values are considered. We solve, for different parameter values, the equilibrium of the airlines’ untolled competition, derive the welfare-maximizing setting and study whether pricing is enough to achieve this setting as an equilibrium. For each parameter, we study an increase and a decrease of 25% of its value, and show three figures that summarize the results. The first of the three figures, to be compared with Figure 6.2, shows the equilibrium in route structures of the untolled competition. If, as a result of a change in a parameter value, the lines that divide the regions in which different equilibria arise move to the left, we conclude that the hub-and-spoke route structure is favored. Phrased differently, when the region in which \((H_x,H_y)\) is the unique equilibrium is larger, we conclude that the parameter change favors the adoption of hub-and-spoke route structures. It follows that if the lines move to the right, i.e. the region in which \((F,F)\) is the unique equilibrium is larger, the parameter variation favors the fully connected route structure.

The second figure, to be compared with Figure 6.3, shows the welfare-maximizing route structure configuration. As a result of our sensitivity analysis, we find that the effect of a change in a parameter value on the welfare-maximizing route structures has the same pattern as for the firms’ choice. This is, whenever a parameter variation enlarges the region where a route structure configuration is the unique equilibrium from the untolled
6.C Sensitivity analysis

competition, it also enlarges the region where that configuration maximizes welfare. The only exception is the value of travel time, where the optimal route structure configuration does not vary significantly in the studied parameter region for $\alpha$.

Finally, the third figure that we show for each parameter change, to be compared with Figure 6.4b, displays the subgame-perfect Nash equilibrium in terms of route structure. This is, the equilibrium in route structure when the regulator applies optimal pricing based on the welfare-maximizing route structure. We find that in all cases, there is a region where $(F, F)$ is an equilibrium of the tolled competition whereas $(H_x, H_y)$ maximizes welfare. Therefore, also when different parameter values are assumed to apply, we find that optimal pricing cannot always decentralize the first-best also in terms of route structure, and an instrument directly aimed at regulating route structure choice may be needed to maximize welfare.

6.C.1 Sensitivity with respect to the demand intercept

Figures 6.8 and 6.9 show that a higher value of the demand intercept favors the fully connected route structures, and that a lower value favors the hub-and-spoke route structures. This is in line with the intuition provided in the previous sections, as higher demand levels favor the fully connected route structures because the losses in connecting markets are larger.

![Figure 6.8: Untolled equilibrium (left), welfare-maximizing configuration (center) and equilibrium under optimal airport pricing (right) in terms of route structure. Higher demand intercept ($A = 5500$).](image)

6.C.2 Sensitivity with respect to the disutility of a unit of schedule delay.

Figures 6.10 and 6.11 reveal that when frequency benefits become more important, through an increase in the disutility of the schedule delay, hub-and-spoke route structures are favored. This is in line with the results in previous literature (e.g., Brueckner, 2004), as hub-and-spoke route structures induce higher frequency in the hub markets.
Chapter 6  Airline route structure competition and network policy

Figure 6.9: Untolled equilibrium (left), welfare-maximizing configuration (center) and equilibrium under optimal airport pricing (right) in terms of route structure. Lower demand intercept ($A = 3300$).

Figure 6.10: Untolled equilibrium (left), welfare-maximizing configuration (center) and equilibrium under optimal airport pricing (right) in terms of route structure. Higher disutility of a unit of schedule delay ($\gamma = 100$).

Figure 6.11: Untolled equilibrium (left), welfare-maximizing configuration (center) and equilibrium under optimal airport pricing (right) in terms of route structure. Lower disutility of a unit of schedule delay ($\gamma = 60$).
6.C.3 Sensitivity with respect to the constant cost per flight.

Figures 6.12 and 6.13 show that higher values of the fixed cost per flight favor the hub-and-spoke route structures. This follows straightforwardly from that the cost benefits that hub-and-spoke networks bring, through a reduced total number of flights, are proportional to the cost per flight (see Eq. (6.19)).

Figure 6.12: Untolled equilibrium (left), welfare-maximizing configuration (center) and equilibrium under optimal airport pricing (right) in terms of route structure. Higher constant cost per flight ($c_f = 21838.5$).

Figure 6.13: Untolled equilibrium (left), welfare-maximizing configuration (center) and equilibrium under optimal airport pricing (right) in terms of route structure. Lower constant cost per flight ($c_f = 13103.1$).
Chapter 6  Airline route structure competition and network policy

6.C.4  Sensitivity with respect to the constant cost per seat.

Figures 6.14 and 6.15 display the result that a higher value of the constant cost per seat favors the fully connected route structure. When the marginal cost of a seat is higher, the connecting passengers’ face higher costs as they require a seat in two legs. As a consequence, the profit that can be made in connecting markets is lower and the social cost of a connection is higher. This is the reason why a higher value of $c_q$ favors the fully connected route structure.

Figure 6.14: Untolled equilibrium (left), welfare-maximizing configuration (center) and equilibrium under optimal airport pricing (right) in terms of route structure. Higher constant cost per seat ($c_q = 103.5$).

Figure 6.15: Untolled equilibrium (left), welfare-maximizing configuration (center) and equilibrium under optimal airport pricing (right) in terms of route structure. Lower constant cost per seat ($c_q = 62.1$).
6.C.5 Sensitivity with respect to the value of travel time.

The results below show that a higher value of travel time favors the adoption of fully connected route structures by (unregulated) firms. As hub-and-spoke route structures induce higher frequencies in the hub markets, they yield increased congestion, which explains why a higher value of $\alpha$ favors the fully connected route structures. This is the opposite of what happens when the disutility of the schedule delay increases. Figures 6.16 and 6.17 show that the optimality of the different route structure configurations is not very sensitive to the value of travel time, at least in the studied range. Finally, recall that a change in the airports’ capacity would have the same effect as a change in the value of time because what matters is the ratio $\alpha/K$ (see Eq. (6.35)). Therefore, there is no need to also vary the value of $K$.

Figure 6.16: Untolled equilibrium (left), welfare-maximizing configuration (center) and equilibrium under optimal airport pricing (right) in terms of route structure. Higher value of travel time ($\alpha = 160.31$).

Figure 6.17: Untolled equilibrium (left), welfare-maximizing configuration (center) and equilibrium under optimal airport pricing (right) in terms of route structure. Lower value of travel time ($\alpha = 96.19$).
Chapter 6  Airline route structure competition and network policy

6.C.6  Sensitivity with respect to the constant transfer cost.

Figures 6.18 and 6.19 display the result that a higher value of the constant transfer cost favors the fully connected route structure. This is a straightforward result: when transferring is more costly for passengers, hub-and-spoke route structures become less attractive for the firm and for welfare maximization.

Figure 6.18: Untolled equilibrium (left), welfare-maximizing configuration (center) and equilibrium under optimal airport pricing (right) in terms of route structure. Higher constant cost transfer cost ($\mu = 50$).

Figure 6.19: Untolled equilibrium (left), welfare-maximizing configuration (center) and equilibrium under optimal airport pricing (right) in terms of route structure. Lower constant cost transfer cost ($\mu = 30$).
Chapter 7

Conclusions
Chapter 7 Conclusions

7.1 Summary of results and implications

This thesis has proposed different theoretical models to answer various research questions concerned with airport pricing policies. It has cast light on the guidelines that airport regulators should follow in order to maximize social welfare by means of setting charges to airlines. This thesis has also studied complementary policies such as network regulations and bans on airport price discrimination. It has applied numerically some of these models, despite their stylized structure, to complement the intuition provided by the analytical study and to shed light on the impact of the analytical results by using parameters that are as much as possible calibrated so as to reflect realistic values.

This thesis has emphasized the role of the airlines’ decisions in terms of fleet timing, which comprises a trade-off between queuing delays and schedule delays, in shaping the socially optimal pricing schemes. It has investigated the policy implications of adopting different assumptions on airline market conduct, such as the models of competition introduced by Cournot (1838), Bertrand (1883) and von Stackelberg (1934). In addition, it has characterized the role of different pricing instruments in achieving the efficient outcome such as time-variant, time-invariant, per-flight and per-passenger airport charges.

Specifically, chapter 2 analyzes airport congestion pricing using Vickrey’s (1969) bottleneck model of congestion at an airport dominated by a single firm. By focusing on the competition between a Stackelberg leader and a competitive fringe, the chapter establishes the relation between the degree of internalization of congestion in the untolled equilibrium with the demand substitution pattern between leader and fringe. This extends the results of previous analysis in static models of congestion, notably those of Brueckner and Van Dender (2008), to a dynamic setting. Importantly, and what is possibly the main conclusion of the chapter, there are multiple pricing schemes that correct the congestion inefficiency and lead to efficient delay levels, and none of them depends crucially on the degree of internalization in the untolled equilibrium. This results from the dynamic nature of the model: the leader’s response when facing an optimally tolled competitive fringe is fully efficient in terms of the timing of the flights, but not in terms of the aggregate number of flights. Therefore, a time-invariant toll can induce the leader to behave efficiently. Yet, time-variant tolls charged to the leader do not cause a distortion in the timing of flights, so that also this charge is able to induce the welfare maximizing outcome.

The implications of the chapter are manifold. The analysis shows that a regulator concerned with pricing an airport that is well represented by the interaction between a dominant airline and a number of small competitors has a diverse set of optimal charges that achieve the social optimum. The different charges vary in the degree of differentiation between leader and fringe and in the implied revenue. Consequently, the regulator can choose the pricing regime that suits better its objectives and possibly enhance the implementation feasibility. Chapter 2 has shown that policy implications obtained from static models can be significantly different from the implications produced by dynamic models. Therefore, in airports where a dynamic model of congestion is more pertinent because, for example, the operational conditions for arrivals (or departures) follow the first-in first-out (FIFO) discipline, caution is required to transfer result and policy recommendations from static models. This FIFO discipline applies, for example, at airports that do not have a
7.1 Summary of results and implications

Chapter 3 extends chapter 2 by studying the equilibrium of a game where firms simultaneously schedule their vehicles. It shows that when firms are homogeneous, a pure strategy Nash equilibrium (PSNE) may not exist and in the cases where it does exist, there are no travel delays. The implications of these results are straightforward. First, in the standard bottleneck model with users that control a single vehicle, a PSNE exists under relative general assumptions (see Lindsey (2004)). Thus, simple congestion-prone systems with firms that are large players can exhibit fundamentally different dynamic behavior than with small (road) users. This, again, has important implications to the transferability of policy recommendations from static models. The chapter also demonstrates that when a PSNE exists, homogeneous firms may incur appreciably different costs even when the outcome is socially efficient. This implies that equity of access to a bottleneck can be an issue when firms control large shares of traffic at congestible facilities. Thus, a regulator concerned with equity may want to regulate the market even though the outcome in terms of timing of the vehicles is efficient.

The chapter finds that when firms differ in their desired arrival time, multiple PSNE may exist in which also no queuing occurs. However, the timing of departures may be inefficient in that departures may begin earlier or later than the social optimum. This result implies that even though there is no queuing, there are still negative externalities of scheduling and there is room for regulation in the market. Nevertheless, the chapter shows that a large share of the potential benefits from regulation are realized in equilibrium due to the internalization of self-imposed congestion by firms. This indicates that a pricing policy in this case would not bring large benefits and if the implementation costs are significant it may be that a laisse-faire policy is socially optimal. The results of this chapter have important modeling implications too. The potential non-existence of equilibrium and the lack of congestion in the equilibria are somewhat puzzling results. Moreover, they do not allow for providing policy recommendations in those cases, which highlights the need for further research in dynamic models. It may be that under different conditions and with additional modeling assumptions equilibrium can be rekindled in a game of simultaneous competition in Vickrey’s (1969) model.

Chapter 4 concentrates on analyzing the internalization of self-imposed congestion under different assumptions on airlines conduct, namely under a Cournot and Bertrand differentiated oligopoly and a duopoly in which one firm is a Stackelberg leader. Under a Bertrand oligopoly, carriers act as if they internalize less congestion than what they impose on themselves, in contrast with a Cournot oligopoly. The chapter finds that the degree of internalization depends on demand–structure parameters, namely the degree of substitutability between the firms’ products, and approaches zero as firms’ products are closer substitutes. This has important implications: when Bertrand competition is pertinent, the welfare gains of congestion pricing are larger, the socially optimal charges are less differentiated, and the degree of self-financing of optimally priced airport infrastructure is higher than under a Cournot oligopoly. Therefore, not only the efficiency gain from congestion pricing is enhanced, but arguably its implementation feasibility as well.

In addition, chapter 4 characterizes the unique role of per-passenger and per-flight charges in achieving the efficient outcome. An important conclusion that can be drawn from the chapter is that both pricing instruments are needed for social welfare maxi-
mization, as congestion externalities need to be addressed through per-flight tolls and the market power inefficiency must be corrected with per-passenger subsidies. There are implications for second-best policies as well. As the two counteracting effects (market power and congestion) are not merged in one pricing instrument, whenever it is optimal to give subsidies and these are not feasible, it is not necessarily the case that the second-best policy is to abstain from charging airlines. The analysis shows that there is large scope for implementing (second-best) congestion pricing when subsidies are not feasible and that its efficiency relative to the first-best is higher, the closer substitute airlines are, and the more firms participate in the market. The numerical analyses suggest that implementing congestion pricing as if airlines would not internalize any congestion may yield substantial benefits, importantly, in the presence of market power distortions.

Chapter 5 investigates airport pricing policies from a different, yet complementary, point of view. It studies under which conditions it is efficient to enforce a ban on price discrimination to airports, for different ownership forms. The chapter shows benefits from allowing price discrimination by a private facility as well as by a public facility. It then compares and characterizes the cases in which enforcing a ban on price discrimination is an efficient policy in terms of social welfare for both types of facilities. A main conclusion of the chapter is that there is a limited scope for the ban to be efficient in both cases. Therefore, an important policy guideline that follows from the analysis is that a broad ban on input price discrimination, such as the European wide ban enforced by the EU directive on airport charges, may have to be revised. It would be more desirable to have a case by case assessment, but, if it is not possible do so, it is not straightforward that enforcing the ban is more efficient than allowing price discrimination.

The result that a ban on price discrimination may be efficient when applied to a public facility but not when applied to a private facility has important policy implications. The ownership form of airports, especially in Europe, has been gradually changing from fully public to private. For example, in 2010, 48% of all European air traffic was handled by a fully or partially private airport (ACI–Europe, 2010). Therefore, the policy of banning price discrimination needs to be particularly reassessed in places where privatization of congestible facilities is increasing.

Chapter 6 deals with studying socially optimal airport pricing policies from a long-run perspective. It investigates how airlines choose their route structure in a given network and compares it with the route structure configuration that maximizes social welfare. The main focus of the chapter is to examine whether the efficient network configuration can arise as an equilibrium when a regulator sets per-passenger and per-flight charges in the network. A main policy conclusion is that a regulator setting airport tolls concerned with correcting output inefficiencies may not achieve the efficient outcome in terms of route structure configuration. Thus, a regulator may benefit from using an additional instrument that is concerned with correcting the network configuration inefficiency directly.

Another main policy implication from the analysis is that information such as marginal benefits and marginal costs is not always enough to design a regulatory scheme that can achieve the social optimum. This is because discrete changes in welfare due to discrete changes in route structure configurations need to be assessed. It is possible that in order to fully reap the long-run benefits from congestion pricing, the tolls should be designed for a different route structure configuration than the one that is observed. The numerical results
suggest that, although the “wrong” network configuration may arise if the regulator does not control directly the route structure choice or cannot assess discrete changes in welfare, the relative welfare losses are limited. So, while the conceptual point that optimal pricing may not decentralize the choice of optimal route structures is significant, the consequent welfare losses may be limited. The latter result, in itself, is of course also an important insight for practical policy making. In any case, it appears to be so in the small numerical example we have developed and its real relevance is a matter of empirical investigation.

In summary, and at the risk of generalizing, the main policy implications from this thesis can be summarized as follows. First, per-flight and per-passenger airport charges are complementary pricing instruments in achieving an efficient outcome from a social welfare standpoint. There is limited substitutability between those instruments and policies aimed at a containment of negative externalities while recognizing the presence of market power should take into account the unique role that each pricing instrument has in achieving the different objectives. Second, in airports where the operational conditions for arrivals or departures follow the first-in first-out discipline, regulators may have a wide set of pricing policies that lead to equally efficient outcomes. This could arguably enhance the implementation feasibility of a socially optimal policy. Third, also in the cases where dynamic congestion models are more pertinent, the outcome of competition between firms may not exhibit congestion and still be inefficient because of schedule delay externalities. This means that regulation in terms of timing of flights in uncongested markets may be beneficial. In addition, even if the outcome is socially efficient, firms may perceive significantly different costs. A policy maker concerned with equity of users (firms and consumers) may want to intervene the market despite its efficiency. Fourth, the airlines’ degree of internalization of self-imposed congestion and thus the extent to which congestion charges set by the airport can increase welfare depend crucially on the market structure and, particularly, on whether Cournot or Bertrand competition is more pertinent. Moreover, under Bertrand competition, the degree of substitutability between the airlines’ products crucially determines the degree of internalization of self-imposed congestion. Finally, broad regulations on price discrimination that extend over numerous transport facilities with different congestion levels and ownership forms may be inefficient.

7.2 Suggestions for future work

This thesis has answered various research questions using different approaches and models. Each of these has, of course, some limitations and naturally raises new research questions, which are discussed in each chapter individually. Nevertheless, we provide below some of the avenues for future research that we deem most important for policy in air transportation and also those that, we believe, have highest scientific relevance.

The analysis in this thesis is mainly theoretical. Many implications and policy recommendations derived from the analyses depend crucially on the magnitude of some effects, which need to be assessed empirically. Empirical analysis of the factors that determine the different conclusions is a natural and an important avenue for future research. An example of this is that congestion pricing when the airlines behave as Bertrand oligopolists brings higher welfare gains than when they act as Cournot oligopolists. This highlights the need
Chapter 7 Conclusions

for empirical research on air transport market conduct. Also the conditions that make a ban on price discrimination to airports with different ownership forms desirable depends crucially on market specific demand and cost characteristics. A thorough investigation of whether and when those conditions hold in places where the ban is implemented is a relevant future research topic.

Chapters 2 and 3 show that systems that are prone to congestion can exhibit fundamentally different behavior when the relevant congestion technology is dynamic rather than static. However, the analysis has been done with a congestion technology that takes the form as assumed in the workhorse dynamic bottleneck model of congestion proposed by Vickrey (1969). Moreover, as chapter 3 shows, simple game settings that are relevant in practice do not have equilibrium with that model. Consequently, theoretical analysis of existence and properties of equilibrium in simultaneous games of competition using different models of dynamic congestion is worth exploring. Examples of such models include the ones proposed by Agnew (1977), Chu (1995) and Mun (1999). These analyses could provide valuable insight into how time-variant airport congestion charges should be implemented in the case of simultaneous competition.

Throughout this thesis the attention has been limited to games of complete information with a fixed number of players. This has allowed for obtaining useful insights into how to design airport pricing policies, but by no means it has allowed for providing the last word on the subject. Further theoretical research would gain from incorporating uncertainty at many levels. Topics include uncertainty of demand and uncertainty of the regulator on airport and airlines operations, notably their cost structure. Investigating the effect of each of the studied pricing policies on the entry of firms and opening of new markets is an important extension of the present analysis.

In most of the analysis of the thesis only restricted attention has been paid to heterogeneity. While some chapters consider heterogeneous firms and others take into account consumer differences between markets, there is still a large unexplored research area regarding heterogeneity. Among the many possibilities for extending the investigation, studying the role of having different types of passengers within markets and airlines applying price discrimination seem to be of particular relevance. Types of heterogeneity that can be incorporated into the analysis include different passengers’ willingness to pay to reduce travel and schedule delays, and different preferred arrival and departure times. The thesis has also assumed that consumers are perfectly informed and are completely rational in their choices. The analyses could be developed further by including more detailed models of consumer behavior, such as modeling the choice of traveling with a firm as a stochastic discrete choice.

Chapter 6 studies the interaction between network effects and optimal pricing policies using a stylized model that considers a small network and that allows for providing valuable insights. However, further analysis is needed to obtain insights into understanding the industrial organization of transport markets in large networks, the process of network formation and the characteristics of equilibria, and the effects that these have on the design of policies aimed at reducing the cost of negative externalities. This is patently an important avenue for future research in terms of policy relevance but also for scientific relevance, as new network models and equilibrium concepts would be required for the analysis.
The thesis has focused on a sequential and non-cooperative relationship between airports and airlines. Further research would benefit from incorporating airport-airline interactions such as agreements, contracts, bargaining and common ownership. The vertical relation that these may have could have important implications for aviation policies.

Finally, a comparison between the efficiency of airport pricing and slot management policies is a relevant avenue for future research. While some studies have already cast light on the issue, the effects of market power, inefficient foreclosure and other potential anticompetitive effects have not been fully incorporated yet. Determining which are the key elements that make one policy to be preferred over the other is a worthwhile avenue for future investigation.
Bibliography


BIBLIOGRAPHY


BIBLIOGRAPHY


Samenvatting (summary in Dutch)


Inmenging van de overheid in deze markt wordt deels gemotiveerd doordat congestie en vertragingen een uitnodiging zijn van negatieve externaliteiten. Deze externaliteiten worden veroorzaakt doordat er kosten zijn van vliegen die niet worden betaald door de passagier, en kosten van congestie die niet worden betaald door de desbetreffende luchtvaartmaatschappij. Een extra vlucht of passagier verhoogt de reistijd, en de daar bijhorende kosten, voor andere vluchten en passagiers door interacties op de start- en landingsbaan, in het luchtruim, in de security zones, of op de gates. Deze externaliteiten worden congestie externaliteiten genoemd. Er zijn ook andere negatieve externaliteiten in deze markt: geluidsoverlast, luchtverontreiniging, en de uitstoot van broeikasgassen zijn de belangrijkste voorbeelden. Sinds Pigou (1920) en Knight (1924) deze zogenoemde congestie-externaliteiten hebben benoemd voor het wegverkeer, weten we dat deze vorm van marktfalen leidt tot niet optimale marktewichten waar, in het niet gereguleerde evenwicht, consumptie (het aantal trips) hoger is dan optimaal.

In de literatuur worden mogelijke oplossingen voor de toenemende congestie op luchthavens intensief bediscussieerd. De belangrijkste alternatieven zijn: het uitbreiden van de capaciteit, het sturen op het totaal aantal vluchtbewegingen via slotcoördinatie, en het invoeren van congestieheffingen. Dit proefschrift behandelt voornamelijk dit laatste alternatief. Het proefschrift heeft tot doel om, vanuit welvaarstperspectief, optimale luchthavenheffingen te bestuderen rekening houdend met congestie-externaliteiten en uitoefening van marktmacht door luchtvaartmaatschappijen. Met andere woorden, wat is de rol van prijsbeleid om congestie aan te pakken als luchtvaartmaatschappijen geen prijnsnemers zijn en grote partijen de mogelijkheid hebben om prijzen en congestie te beïnvloeden? Om efficiënt prijsbeleid te onderzoeken, modelleren dit proefschrift zowel het keuzegedrag van reizigers als het strategisch gedrag van luchtvaartmaatschappijen. Aan de hand van verschillende theorethische modellen worden uiteenlopende onderzoeksvragen beantwoord. Het onderzoek geeft richtlijnen voor overheden en beleidsmakers over hoe congestie-heffingen in de luchtvaart kunnen bijdragen aan het maximaliseren
van welvaart. Naast heffingen worden ook andere beleidsopties besproken, zoals het direct reguleren van de netwerkkeuze van luchtvaartmaatschappijen en het al dan niet toestaan van prijsdiscriminatie op luchthavens. De numerieke analyses in het proefschrift helpen het begrip van de theoretische resultaten te verscherpen, waarbij de kalibratie van deze numerieke modellen zo goed mogelijk de werkelijke markten beschrijft.

Dit proefschrift onderstreep de rol die luchtvaartmaatschappijen hebben bij het tot stand komen van optimale heffingen: via het opstellen van vluchtschema’s, waarin elke keuze een afweging is tussen vertraging en congestie of een afwijkings van de optimale geplande aankomst- en vertrekken, beïnvloeden luchtvaartmaatschappijen dit proces. Dit proefschrift beschrijft ook de beleidsimplicaties van het gebruik van verschillende veronderstellingen over de structuur van de luchtvaartmarkt, zoals Cournot (1838), Bertrand (1883) of von Stackelberg (1934) concurrentie, en de rol van verschillende prijsinstrumenten om het welvaarts optimum te bereiken, zoals tijdsafhankelijke, tijdsonafhankelijke, heffingen per vlucht en per passagier.

Hoofdstuk 2 analyseert congestieheffingen op een door één luchtvaartmaatschappij gedomineerde luchthaven aan de hand van Vickrey’s (1969) knelpuntmodel van congestie: het zogenaamde bottleneckmodel. In tegenstelling tot de bestaande literatuur combineert dit hoofdstuk een dynamisch congestiemodel en een verticaal, multi-level concurrentiemodel waarin het gedrag van luchtvaartmaatschappijen en passagiers expliciet wordt meegenomen. In een model met een Stackelberg leider en kleine prijswinners (de competitive fringe), laat dit hoofdstuk zien dat er een relatie is tussen hoeveel congestie wordt geïnternaliseerd in de niet gereguleerde markt, en de mate van vraagsubstitutie tussen de leider en de fringe. Deze analyse breidt de bestaande resultaten op basis van statische congestiemodellen (zie Brueckner and Van Dender, 2008) uit naar een dynamische setting. Een belangrijke conclusie is dat er meerdere prijsschema’s mogelijk zijn die resulteren in een optimaal congestieniveau, waaronder schema’s met en zonder tijdsafhankelijke heffingen. Deze schema’s worden niet fundamenteel beïnvloed door de mate waarin congestie wordt geïnternaliseerd in het niet gereguleerde evenwicht. Het dynamische karakter van het model verklaart deze resultaten: het gedrag van de leider die wordt geconfronteerd met een optimaal beprijsde fringe is volledig efficiënt in het bepalen van het vluchtschema (aankomst- en vertrekken), maar niet in het totaal aantal vluchten. Hierdoor kan een niet-tijdsafhankelijke heffing de leider al bewegen tot efficiënt gedrag, namelijk het aanpassen van het aantal vluchten. Tijdsafhankelijke heffingen leiden niet tot een verstoring in het opstellen van vluchtschema’s, waardoor deze heffingen ook resulteren in het welvaarts optimum.

Dit hoofdstuk biedt meerdere nieuwe beleidsinzichten. De analyse laat zien dat beleidsmakers een diverse set van optimale heffingen kunnen hanteren om het welvaarts optimum te bereiken op luchthavens met één dominante luchtvaartmaatschappij en meerdere kleinere concurrenten. Deze verschillende schema’s van heffingen verschillen zowel in de mate van differentiatie tussen de leider en de fringe-maatschappijen, als in de resulterende heffingsopbrengst voor de overheid. De regulator kan dus het schema van heffingen kiezen dat past bij haar doelstellingen en daarmee, mogelijkerwijs, de implementatietereenvoudigen. Hoofdstuk 2 laat verder zien dat de beleidsimplicaties die worden verkregen uit statische modellen significant anders kunnen zijn dan die welke volgen uit dynamische modellen. Terughoudendheid is van belang op het moment dat resultaten en
beleidsinzichten gebaseerd op statische modellen worden toegepast op luchthavens waar congestie een duidelijk dynamisch karakter heeft, bijvoorbeeld doordat aankomende (of vertrekkende) vluchten op een first-in first-out (FIFO) basis worden afgehandeld. Dit FIFO principe geldt bijvoorbeeld op luchthavens zonder slotcoördinatie.

Hoofdstuk 3 is een uitbreiding op hoofdstuk 2, en kijkt naar de simultane beslissingen van luchtvaartmaatschappijen betreffende hun vluchtschema’s. Dit hoofdstuk analyseert de eigenschappen van het evenwicht in het Vickrey bottleneckmodel in het geval dat elk bedrijf een substantieel deel van de totale vraag bedient. Bedrijven kiezen simultaan hun vertrekschema’s voor hun eigen vloot en elk bedrijf internaliseert hierbij de congestie die elk van zijn vliegtuigen veroorzaakt op alle andere vliegtuigen. De analyse laat zien dat als bedrijven homogeen zijn, een Nash evenwicht in pure strategieën niet hoeft te bestaan. Als het evenwicht wél bestaat, is het een evenwicht waarin geen vertragingen bestaan. In het standaard bottleneckmodel met atomistische gebruikers, is het bestaan van een Nash evenwicht in pure strategieën aan te tonen onder relatief algemene veronderstellingen (zie Lindsey, 2004). Met andere woorden, in markten met congestie en grote bedrijven is het dynamisch gedrag fundamenteel anders dan in markten met atomistische gebruikers. Dit laat nogmaals zien dat beleidsaanbevelingen rekening moeten houden met zowel de dynamische context als met de marktstructuur. Dit hoofdstuk toont aan dat wanneer een Nash evenwicht in pure strategieën bestaat, homogene bedrijven substantieel andere kosten kunnen hebben, zelfs in het welvaartsoptimum. Er is dus sprake van een verdelingsvraagstuk rond de toegang tot en het gebruik van de bottleneck als bedrijven marktmacht hebben en grote delen van de vraag bedienen. Zelfs als de vluchtschema’s efficiënt zijn vanuit een welvaartsperspectief, kan dit verdelingsvraagstuk een reden zijn om de markt te reguleren.

Er kunnen meerdere Nash evenwichten zonder vertraging bestaan als bedrijven verschillende gewenste aankomsttijden hebben. De vluchten kunnen hierdoor eerder of later vertrekken dan sociaal optimaal is. Dus zelfs zonder vertraging in het evenwicht zijn er negatieve externaliteiten via de vluchtschema’s en kan regulatie dus wenselijk zijn. Een groot gedeelte van de potentiële baten van het reguleren wordt echter al gerealiseerd door het internaliseren van de congestie-effecten op de eigen vloot. Het beprijzen van de congestie zal dus niet leiden tot hoge additionele baten; sterker nog, gelet op de kosten om een dergelijk beleid te implementeren, kan het volgens dit model zelfs zo zijn dat een laissez-faire beleid optimaal is.

Hoofdstuk 4 analyseert hoe luchtvaartmaatschappijen congestie internaliseren onder verschillende marktstructuren met gedifferentieerde producten: Cournot oligopolie, Bertrand oligopolie, en een Stackelberg duopolie model. In een Cournot oligopolie internaliseert een vliegmaatschappij de congestie die zij op haar eigen vliegtuigen oplegt volledig. Echter, in een Bertrand oligopolie internaliseren vliegmaatschappijen een kleiner deel van de congestie. De mate waarin eigen congestie wordt geïnternaliseerd hangt af van de substitueerbaarheid tussen de producten van de verschillende bedrijven, en daalt richting nul als de producten nauwere substituten worden. Dit houdt in dat als de markt gekenmerkt wordt door Bertrand in plaats van Cournot concurrentie de heffingen minder neerwaarts hoeven te worden aangepast vanwege zelf-internalisatie. De welvaartsverbeteringen van het beprijzen van congestie zijn dan groter; de bijbehorende optimale heffingen zijn minder gedifferentieerd; en de mate van zelf-financiering van luchthaven
infrastructuur is ook groter. Niet alleen is de efficiency van reguleren dus groter, het is ook mogelijk dat het implementeren ervan beter haalbaar is.

Hoofdstuk 4 beschrijft daarnaast de rol van heffingen per passagier en per vlucht. Een belangrijke conclusie is dat beide heffingen noodzakelijk zijn voor welvaartsmaximalisatie: de congestie-externaliteiten worden via de heffingen per vlucht beïnvloed, en marktmacht via (negatieve) heffingen per passagier (subsidies dus). Omdat de twee tegengestelde effecten (congestie en marktmacht) niet in één prijsinstrument zijn verenigd, kunnen in het second-best beleid—als subsidies niet mogelijk zijn, maar wel optimaal vanuit een welvaartsperspectief—dus alsnog heffingen worden ingevoerd voor de luchtvaartmaatschappijen. Uit de analyse blijkt dat er mogelijkheden zijn voor het implementeren van (second-best) congestie heffingen als subsidies niet mogelijk zijn. De efficiency van dit second-best beleid is groter naarmate de luchtvaartmaatschappijen betere substituten zijn en dus minder congestie zelf internaliseren, en er meer bedrijven in de markt actief zijn. De numeriek analyses suggereren dat het invoeren van congestieheffingen, onder de aanname dat luchtvaartmaatschappijen geen enkele congestie kosten internaliseren, hoge baten genereert als er daadwerkelijke verstoringen zijn als gevolg van marktmacht.

Hoofdstuk 5 biedt een analyse van het prijsbeleid voor luchthavens vanuit een andere, maar complementaire, dimensie. Dit hoofdstuk bestudeert onder welke voorwaarden het efficiënt is om luchthavens te verbieden prijsdiscriminatie toe te passen, rekening houdend met de verschillende eigendomsstructuren van luchthavens. Input prijsdiscriminatie door een private luchthaven, oftewel verschillende prijzen hanteren voor verschillende luchtvaartmaatschappijen, kan zowel de totale productie als de welvaart laten toenemen. In een vergelijkbare situatie zonder congestie, zou de geaggregeerde output niet veranderen, en de welvaart dalen als gevolg van input prijsdiscriminatie. Het bestaan van negatieve consumptie-externaliteiten maakt input prijsdiscriminatie dus meer wenselijk. Dit hoofdstuk laat ook de baten zien van prijsdiscriminatie op een luchthaven in publiek eigendom, en beschrijft onder welke voorwaarden het verbieden van prijsdiscriminatie efficiënt is voor publieke en private luchthavens. Een belangrijke conclusie is dat in veel gevallen het verbieden van prijsdiscriminatie niet wenselijk is vanuit het welvaartsperspectief. Dat geldt zowel voor publieke als private luchthavens. Hieruit volgt de beleidsaanbeveling dat een breed verbod op input prijsdiscriminatie, zoals door Europese regelgeving uitgevaardigd voor luchtaventarieven in heel Europa, wellicht moet worden herzien. De eigendomsstructuur van luchthavens, vooral in Europa, is geleidelijk veranderd van publiek naar privaat eigendom. In 2010, bijvoorbeeld, werd 48% van al het Europese luchtverkeer afgehandeld door volledig of gedeeltelijk geprivatiseerde luchthavens ACI–Europe (2010). In het bijzonder daar waar de privatisering van publieke voorzieningen verder toeeneemt, moet het beleid rondom het verbod op prijsdiscriminatie worden herzien.

Hoofdstuk 6 bestudeert optimale luchthavencongestie heffingen vanuit het lange termijn perspectief. In dit hoofdstuk wordt de keuze die luchtvaartmaatschappijen maken ten aanzien van hun routestructuur vergeleken met de routestructuur die welvaart zou maximaliseren. De focus ligt op de vraag of in het evenwicht een efficiënte routestructuur kan ontstaan door heffingen per passagier en vlucht in te voeren. Een belangrijke uitkomst is dat luchthavenheffingen bedoeld om output inefficiënties te corrigeren niet noodzakelijk leiden tot een efficiënt evenwicht in de route structuur. Dit betekent dat een regulator dus een derde beleidsinstrument nodig heeft om de route structuur direct te reguleren.
Een andere belangrijke beleidsimplicatie is dat de informatie over de marginale baten en marginale kosten niet altijd genoeg is om het welvaartsoptimum te kunnen bereiken. De reden hiervoor is dat de discrete veranderingen in routestructuur leiden tot discrete veranderingen in welvaart. Het is mogelijk dat, om de maximale totale maatschappelijke baten van congestieheffingen op de lange termijn te realiseren, de heffingen bepaald moeten worden voor een andere dan de waargenomen routestructuur. De numerieke analyses laten zien dat, alhoewel het “verkeerde” netwerk kan ontstaan als de regulator de netwerkkeuze niet controleert, het welvaartsverlies van dit niet-optimale netwerk heel beperkt is. Dit hangt samen met het feit dat het verschil tussen het welvaart-optimale netwerk en winst-maximaliserende netwerk beperkt is. Dus, ook al is aparte route-regulering mogelijk nodig om het optimum te bereiken, het welvaartsverlies van dit niet doen is waarschijnlijk beperkt.

De belangrijkste beleidsimplicaties op basis van dit proefschrift kunnen, op basis van het bovenstaande en met het risico te veel te generaliseren, als volgt worden samengevat. Ten eerste, luchthavenheffingen per vlucht en per passagier zijn complementaire prijsinstrumenten. Er is slechts een beperkte substitutie mogelijk tussen deze instrumenten. Beleid gericht op het corrigeren van negatieve externaliteiten in markten waar bedrijven marktmacht uitoefenen, moet rekening houden met de rol die elk van de twee instrumenten speelt bij het bereiken van verschillende doelstellingen. Ten tweede, voor luchthavens waar aankomende (of vertrekkende) vluchten op een first-in first-out (FIFO) basis worden afgehandeld en er een duidelijke leader-fringe markstructuur is, hebben beleidsmakers keuze tussen verschillende vormen van prijsbeleid die elk leiden tot hetzelfde welvaartsoptimum. Ten derde, bij concurrentie tussen luchtvaartmaatschappijen kan het zijn dat er geen evenwicht is in aankomsttijden, of kan het zijn dat er alleen een evenwicht zonder congestie is. Dit evenwicht is dan toch nog niet optimaal omdat er nog steeds externaliteiten met betrekking tot de schedule delay zijn; dit geldt ook in dynamische congestiemodellen. Dit betekent dat het reguleren van de vluchtschema’s in markten zonder congestie welvaart kan vergroten. Zelfs als zonder overheidsingrijpen het evenwicht efficiënt is, kan de ongelijke verdeling van kosten tussen bedrijven een reden zijn voor interventie. Een volgende beleidsconclusie is dat de mate waarin luchtvaartmaatschappijen zelf congestie internaliseren, en daarmee de mate waarin congestieheffingen ingevoerd door de luchthaven welvaartsverhogend zijn, afhanger van de markstructuur en meer in het bijzonder van de vraag van Cournot of Bertrand concurrentie is, en in dat tweede geval hoe substitueerbaar de producten van de verschillende aanbieders zijn. Tenslotte zijn de algemene regelgeving en beperkingen rondom prijsdiscriminatie, welke gelden voor vele transport voorzieningen met verschillende niveaus van congestie en eigendomsstructuren, mogelijk niet efficiënt.
The Tinbergen Institute is the Institute for Economic Research, which was founded in 1987 by the Faculties of Economics and Econometrics of the Erasmus University Rotterdam, University of Amsterdam and VU University Amsterdam. The Institute is named after the late Professor Jan Tinbergen, Dutch Nobel Prize laureate in economics in 1969. The Tinbergen Institute is located in Amsterdam and Rotterdam. The following books recently appeared in the Tinbergen Institute Research Series:

563 P. SUN, Tail Risk of Equidity Returns
564 C.G.J. KARSTEN, The Law and Finance of M&A Contracts
565 C. OZGEN, Impacts of Immigration and Cultural Diversity on Innovation and Economic Growth
566 R.S. SCHOLTE, The interplay between early-life conditions, major events and health later in life
567 B.N. KRAMER, Why don’t they take a card? Essays on the demand for micro health insurance
568 M. KILIÇ, Fundamental Insights in Power Futures Prices
569 A.G.B. DE VRIES, Venture Capital: Relations with the Economy and Intellectual Property
570 E.M.F. VAN DEN BROEK, Keeping up Appearances
571 K.T. MOORE, A Tale of Risk: Essays on Financial Extremes
572 F.T. ZOUTMAN, A Symphony of Redistributive Instruments
573 M.J. GERRITSE, Policy Competition and the Spatial Economy
574 A. OPSCHOOR, Understanding Financial Market Volatility
575 R.R. VAN LOON, Tourism and the Economic Valuation of Cultural Heritage
576 I.L. LYUBIMOV, Essays on Political Economy and Economic Development
577 A.A.F. GERRITSEN, Essays in Optimal Government Policy
578 M.L. SCHOLTUS, The Impact of High-Frequency Trading on Financial Markets
579 E. RAVIV, Forecasting Financial and Macroeconomic Variables: Shrinkage, Dimension reduction, and Aggregation
580 J. TICHEM, Altruism, Conformism, and Incentives in the Workplace
581 E.S. HENDRIKS, Essays in Law and Economics
582 X. SHEN, Essays on Empirical Asset Pricing
583 L.T. GATAREK, Econometric Contributions to Financial Trading, Hedging and Risk Measurement
584 X. LI, Temporary Price Deviation, Limited Attention and Information Acquisition in the Stock Market
585 Y. DAI, Efficiency in Corporate Takeovers
586 S.L. VAN DER STER, Approximate feasibility in real-time scheduling: Speeding up in order to meet deadlines
587 A. SELIM, An Examination of Uncertainty from a Psychological and Economic Viewpoint
588 B.Z. YUESHEN, Frictions in Modern Financial Markets and the Implications for Market Quality
589 D. VAN DOLDER, Game Shows, Gambles, and Economic Behavior
S.P. CEYHAN, Essays on Bayesian Analysis of Time Varying Economic Patterns
S. RENES, Never the Single Measure
D.L. IN ’T VELD, Complex Systems in Financial Economics: Applications to Interbank and Stock Markets
Y.YANG, Laboratory Tests of Theories of Strategic Interaction
M.P. WOJTOWICZ, Pricing Credits Derivatives and Credit Securitization
R.S. SAYAG, Communication and Learning in Decision Making
S.L. BLAUW, Well-to-do or doing well? Empirical studies of wellbeing and development
T.A. MAKAREWICZ, Learning to Forecast: Genetic Algorithms and Experiments
P. ROBALO, Understanding Political Behavior: Essays in Experimental Political Economy
R. ZOUTENBIER, Work Motivation and Incentives in the Public Sector
M.B.W. KOBUS, Economic Studies on Public Facility use
R.J.D. POTTER VAN LOON, Modeling non-standard financial decision making
G. MESTERS, Essays on Nonlinear Panel Time Series Models
S. GUBINS, Information Technologies and Travel
D. KOPÁNYI, Bounded Rationality and Learning in Market Competition
N. MARTYNова, Incentives and Regulation in Banking
D.KARSTANJE, Unraveling Dimensions: Commodity Futures Curves and Equity Liquidity
T.C.A.P. GOSENS, The Value of Recreational Areas in Urban Regions
L.M. MARČ, The Impact of Aid on Total Government Expenditures
C.LI, Hitchhiking on the Road of Decision Making under Uncertainty
L. ROENDAHL HUBER, Entrepreneurship, Teams and Sustainability: a Series of Field Experiments
X. YANG, Essays on High Frequency Financial Econometrics
A.H. VAN DER WEIJDE, The Industrial Organization of Transport Markets: Modeling Pricing, Investment and Regulation in Rail and Road Networks